Journal of Memory and Language 93 (2017) 276–303

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

The multimodal nature of spoken word processing in the visual world: Testing the predictions of alternative models of multimodal integration

Alastair C. Smith [a,*], Padraic Monaghan [b], Falk Huettig [a,c]

[a] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[b] Department of Psychology, Lancaster University, Lancaster, UK
[c] Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 10 May 2015
revision received 23 August 2016

Keywords:
Visual world paradigm
Visual attention
Spoken word recognition
Connectionist modelling
Multimodal processing

ABSTRACT

Ambiguity in natural language is ubiquitous, yet spoken communication is effective due to integration of information carried in the speech signal with information available in the surrounding multimodal landscape. Language mediated visual attention requires visual and linguistic information integration and has thus been used to examine properties of the architecture supporting multimodal processing during spoken language comprehension. In this paper we test predictions generated by alternative models of this multimodal system. A model (TRACE) in which multimodal information is combined at the point of the lexical representations of words generated predictions of a stronger effect of phonological rhyme relative to semantic and visual information on gaze behaviour, whereas a model in which sub-lexical information can interact across modalities (MIM) predicted a greater influence of visual and semantic information, compared to phonological rhyme. Two visual world experiments designed to test these predictions offer support for sub-lexical multimodal interaction during online language processing.

Introduction

One of the defining features of language is displacement, i.e., the fact that concepts need not refer to objects or events that are currently present (Hockett & Altmann, 1968). In line with this observation is a long tradition of research in the language sciences which has largely ignored potential influences of 'non-linguistic' information sources (e.g., Fodor, 1983). However, although language does not need to refer to objects which are physically present it is often used in such a way. Moreover, psycholinguistic research over recent years suggests that language processing (including spoken word processing) is highly interactive in terms of combining multiple information sources to form an interpretation of the signal (see Onnis & Spivey, 2012). It is therefore likely to be a profound misrepresentation to restrict models of spoken word recognition exclusively to auditory information, overlooking multimodal aspects of the speech processing system (e.g. Luce, Goldinger, Auer, & Vitevitch, 2000; McClelland & Elman, 1986; Norris & McQueen, 2008; Scharenborg & Boves, 2010).

Indeed, the prevalence of ambiguity in natural language (Piantadosi, Tily, & Gibson, 2012) is evidence for the efficiency with which the human speech processing system integrates linguistic and extra-linguistic information. If we accept that language usage takes place in context (i.e., embedded within extra-linguistic factors, such as visual

[*] Corresponding author at: Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands.
E-mail address: alastair.smith@mpi.nl (A.C. Smith).

http://dx.doi.org/10.1016/j.jml.2016.08.005
0749-596X/© 2016 Elsevier Inc. All rights reserved.

environment, non-verbal communicative cues, world knowledge, and so on) then the amount of information an efficient language should convey must be less than the amount of information required out of context (Kurumada & Jaeger, 2015; Monaghan, Christiansen, & Fitneva, 2011). However, we know ambiguity in natural language is ubiquitous yet such ambiguity is rarely harmful to effective communication (Ferreira, 2008; Jaeger, 2006, 2010; Piantadosi et al., 2012; Roland, Elman, & Ferreira, 2006; Wasow & Arnold, 2003; Wasow, Perfors, & Beaver, 2005). This implies that the speech processing system is able to efficiently integrate extra-linguistic contextual information with the ambiguous speech stream it receives. The lack of explicit awareness we have of the level of ambiguity within the raw speech signal when processing speech in natural settings illustrates the speed and ease with which linguistic and non-linguistic information is integrated by the human speech processing system.

Models of speech recognition and speech comprehension have frequently overlooked this multimodal aspect of the speech processing system (e.g., Luce et al., 2000; McClelland & Elman, 1986; Norris & McQueen, 2008; Scharenborg & Boves, 2010), with comparatively little known about the architecture that supports integration and the temporal structure of this process. In this study we test two explicit implementations of alternative hypotheses describing how visual, phonological and semantic information may be integrated when processing spoken words in a visual world. The first model is based on TRACE (McClelland & Elman, 1986) and multimodal information integration occurs over lexical representations. The alternative model permits integration of multimodal information over sub-lexical representations. These simulations generate similar predictions for the role of phonologically similar words in competition when the similarity is at the word onset. However, critically, they provide contrasting predictions for the influence of phonological rhyme information on fixation behaviour relative to visual and semantic information during online spoken word processing. We therefore tested these effects in two visual world eye-tracking experiments (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). The results provide constraints on when and how such information is integrated in speech processing.

Models of multimodal integration during speech processing

A distinct division in perspectives continues to exist within both cognitive psychology and cognitive neuroscience regarding the characterisation of how and when non-linguistic and linguistic information interact during speech processing (e.g. Dilkina, McClelland, & Plaut, 2010; Leonard & Chang, 2014; Pulvermüller, Shtyrov, & Hauk, 2009).

The classical view within psycholinguistics argues that on hearing a spoken word information in the speech signal activates progressively larger units of representation within a modular phonological processing hierarchy, for example progressing from activation of primary phonetic features, to phonemes, to ultimately activating lexical units (e.g. McClelland & Elman, 1986). It is at this point, at the lexical level, that information in other modalities can connect to influence processing (e.g. Fodor, 1983; Friederici, 2002; Marslen-Wilson, 1987; Spivey, 2007), although such architectures can vary greatly in the extent to which information is able to interact between levels (see, e.g., McClelland, Mirman, & Holt, 2006; McQueen, Norris, & Cutler, 2006).

Alternatively, information in other modalities may be available to interact sub-lexically (e.g. Dilkina, McClelland, & Plaut, 2008; Dilkina et al., 2010; Gaskell & Marslen-Wilson, 1997; Pulvermüller et al., 2009). In such an architecture it becomes feasible for associations to develop between sub-lexical representations across modalities, for example between individual phonemes and individual semantic features.

In this paper we implement each of these alternative architectures in cognitively plausible (McClelland, Mirman, Bolger, & Khaitan, 2014) computational models. In both cases spoken word recognition and spoken word comprehension are framed in terms of multimodal constraint satisfaction (cf. MacDonald, Pearlmutter, & Seidenberg, 1994; McClelland, Rumelhart, & Hinton, 1986; McClelland et al., 2014), with words conceived as entities that connect representations across multiple modalities (e.g., phonological, orthographic, semantic, visual, etc.). In both models, speech processing occurs in a multimodal context, with activation of information passing between modalities to reflect real time sensory input. Both models are able to incorporate such multimodal cues to adapt their response in accordance to the current information available.

The two models differ however in the level at which multimodal information is able to interact. To represent a lexical level multimodal interaction model we extend the TRACE model of speech processing (McClelland & Elman, 1986) to allow activation cascading from visual and semantic representations to influence processing at the lexical level. TRACE provides a phonological processing hierarchy that allows activation to interact bi-directionally between three levels of representations: phonetic features, phonemes and words. We extend this system by injecting activation from visual and semantic levels into the TRACE hierarchy at the lexical level.

For contrast, we also implement a fully interactive system in which information at all levels of representation is free to combine across modalities. To represent such a system, we use the Multimodal Integration Model (MIM) of language processing which integrates concurrent phonological, semantic and visual information in parallel during spoken word processing (Smith, Monaghan, & Huettig, 2013, 2014a, 2014b; see also Monaghan & Nazir, 2009). The model is derived from the Hub-and-Spoke framework (Dilkina et al., 2008, 2010; Plaut, 2002; Rogers et al., 2004), a single system architecture that consists of a central resource (hub) that integrates and translates information between multiple modality specific sources (spokes). Critically, processing in the MIM is emergent, with minimal assumptions regarding initial connectivity or constraints on the flow of information within the network. Behaviour is thus a consequence of the system learning to map across modalities in which differing representational structures are embedded.

Visual world eye-tracking as a method to study spoken word processing

Visual world experiments, in which participants' gaze is recorded when mapping between visual and auditory stimuli, have been used extensively to examine the interface between visual and linguistic processing streams (see Huettig, Rommers, & Meyer, 2011, for a review). These studies provide insight into the type of information activated as a spoken word unfolds, the relative influence of specific sources of information during speech comprehension, and the temporal structure of this process. Such insights are based on the assumption that gaze towards an item reflects the level to which properties of the item (relative to all other items within the display) are activated at a given point in time by the speech signal (see Ferreira & Tanenhaus, 2007; Huettig, Mishra, & Olivers, 2012; Tanenhaus, Magnuson, Dahan, & Chambers, 2000).

We know from visual world studies that objects in the visual environment whose names share their phonological onset with a spoken target word (e.g., beaver and beaker) can attract visual attention from shortly after word onset (Allopenna, Magnuson, & Tanenhaus, 1998). We also know from the same study that visually displayed objects whose names share their phonological rhyme with the spoken target word (e.g., speaker and beaker) are also fixated more than unrelated objects shortly after target word onset, yet slightly later than objects that share their phonological onsets. But it is not only the activation of phonological information that has been indexed by such studies of language mediated visual attention. They have also demonstrated that items that share visual properties (e.g., shape: beaker and bobbin) with a spoken target word (but no phonological relationship) attract attention early post word onset (Dahan & Tanenhaus, 2005; Huettig & Altmann, 2007; Huettig & McQueen, 2007). Items that share semantic (but not phonological or visual) relationships with spoken target words (e.g., cent and purse) also have been demonstrated to attract attention rapidly post word onset (Duñabeitia, Avilés, Afonso, Scheepers, & Carreiras, 2009; Huettig & Altmann, 2005; Yee, Overton, & Thompson-Schill, 2009; Yee & Sedivy, 2006; Yee, Huffstetler, & Thompson-Schill, 2011). Together, these data demonstrate that as a spoken word unfolds, its phonological, visual and semantic properties are activated rapidly and thus can be recruited to map onto information extracted from the immediate visual environment.

To examine the relative timing of activation of phonological, semantic and visual information by the unfolding speech signal, Huettig and McQueen (2007) presented participants with scenes containing items that shared properties of the target word in one of each of these three dimensions. Scenes contained an item which shared its phonological onset with the spoken target word (phonological onset competitor); an item that shared visual properties with the spoken target word (visual competitor); an item that shared semantic properties with the spoken target word (semantic competitor); and an item that was unrelated to the spoken word in all three dimensions (unrelated distractor). They observed that participants first looked towards phonological competitors while later looking towards visual and semantic competitors once later phonemes had provided disambiguating information to discount the phonological competitor. This pattern of gaze was interpreted by Huettig and McQueen as evidence for the cascaded activation of information through the speech processing system, with the speech signal first activating the target word's phonological properties, then later visual and semantic properties.

Similarly, pairing items within the visual display that contrast in the properties they share with the spoken target word has also been used to examine the relative influence of a given property on language mediated eye gaze, and, by extension, motivate statements regarding relative activation during spoken word processing. Allopenna et al. (1998) presented scenes containing items that either shared their phonological onset or rhyme with the spoken target word. They observed that participants' gaze towards phonological rhyme competitors occurred later and was weaker than onset effects. Studies of rhyme competitor effects have since shown that they typically result in only small, marginally significant effects (see also Allopenna et al., 1998; McQueen & Huettig, 2012; McQueen & Viebahn, 2007). This indicates that phonological information in the onset is more influential in spoken word recognition than information carried in the rhyme.

The use of language mediated eye gaze to make statements about spoken word recognition has gained influence due to a coupling of visual world data and computational models of spoken word recognition. This approach requires the explicit description of the mechanisms driving eye gaze that can be tested against behavioural findings. Allopenna et al.'s (1998) observation of an influence of rhyme competitors on fixation behaviour proved notable as this was initially believed to be a point of distinction between alternative models of spoken word recognition: such as continuous mapping models (e.g. TRACE: McClelland & Elman, 1986) and alignment models (e.g. Marslen-Wilson, 1987; Norris, 1994). In early descriptions of alignment models, initial phonemes constrain the candidate set of words such that words that mismatched at onset, such as rhyme competitors, are no longer under consideration. Hence, should such an alignment model be driving fixation behaviour, then fixation of rhyme competitors should not exceed levels displayed towards unrelated items. In contrast, within continuous mapping models, mapping occurs across the entire word with overall similarity driving a word's level of activation. Thus, words that share their rhyme, yet not their onset, will still be activated. TRACE, the continuous mapping model tested in Allopenna et al. (1998), predicts both a rhyme effect, and also a distinction in the level of activation of onset and rhyme competitors. As onset phonemes are encountered earlier, their activation will, before the overlapping phonemes in the rhyme unfold, inhibit rhyme competitors. Hence, TRACE predicts that rhyme competitors will be activated at levels lower than those of onset competitors, which was the pattern observed in Allopenna et al. (1998). Although continuous mapping models had predicted the influence of phonological rhyme overlap during spoken word recognition, evidence for such an influence had been difficult to isolate using standard priming paradigms (Andruski, Blumstein, & Burton, 1994; Connine, Blasko, & Titone, 1993). Eye gaze in the visual world paradigm, however, offers a temporally rich measure that provided the necessary sensitivity to capture these subtle effects (Allopenna et al., 1998).

It has since been demonstrated that alignment models are also capable of generating rhyme competitor effects if they are exposed to noise in the learning environment, such that onset information is not always a perfect predictor of the target word (Magnuson, Tanenhaus, & Aslin, 2000; Magnuson, Tanenhaus, Aslin, & Dahan, 2003; Smith et al., 2013). Evidence to support such predictions is provided by recent visual world data that demonstrates that onset and rhyme effects on language mediated eye gaze can be modulated by the level of noise participants are exposed to in the speech signal (McQueen & Huettig, 2012).

In sum, studies of language mediated visual attention have demonstrated that visually displayed items that share their phonological rhyme with the spoken target word attract attention more than unrelated items. However, such effects have been small and tend to have only been observed under heavily controlled laboratory conditions, in which phonology is the only property connecting items in the display to the spoken target word. Therefore, it remains an open question whether phonological rhyme information exerts an influence on language mediated eye gaze when other sources of information are available to map between visual and auditory streams, which is a closer simulation of day-to-day spoken word processing, in situations when information from semantic or visual modalities may also be available to constrain spoken word recognition and comprehension.

Aims of the current study

Our aim is to examine the interaction of phonological rhyme, semantic and visual information within language mediated visual attention. The literature outlined above demonstrates that language mediated eye gaze is dependent on the interaction of phonological, visual and semantic information, it therefore offers a novel means of examining how such sources of information may interact when mapping between visual and linguistic streams. These data motivate constraints regarding the architecture supporting such multimodal interaction during spoken word processing. We first test two alternative models, the MIM and extended TRACE model, to generate predictions for how gaze is predicted to be distributed towards visual, semantic and phonological rhyme competitors when visual, semantic and phonological information are integrated at different points in lexical processing. The key distinguishing data between these accounts turns out to be derived from studies of rhyme competitor effects in the visual world paradigm that have not yet been tested experimentally. Therefore, two visual world experiments are then presented to measure behaviour of participants when exposed to the conditions simulated in the models, in order to distinguish between these alternative models.

The first visual world experiment presents scenes that contain a single phonological rhyme competitor and three unrelated distractors. This will establish whether relationships within the materials are sufficient to generate the rhyme effect reported in previous visual world studies. The second visual world experiment presents the same scenes as used in the first experiment but with two of the unrelated distractors replaced with a visual and a semantic competitor, to more closely reflect lexical processing in situations when multimodal information sources are simultaneously available. The second experiment thus examines how the phonological rhyme effect is affected by competition from semantic and visual competitors.

A comparison between Experiment 1 (rhyme competitor only) and Experiment 2 (rhyme, semantic and visual competitors) offers four possible outcomes: (1) the rhyme effects are not altered by the presence of visual and semantic competitors; (2) the rhyme effect is weakened; (3) the rhyme effect increases; or (4) the rhyme effect is eliminated, thus providing an additional means of evaluating model fit. Examining the results of Experiment 2 in isolation also provides a rich data set against which alternative model predictions can be tested for the point of interaction of different information sources in lexical processing. The extent to which alternative models can simulate the observed effects of phonological competitors alongside the influence of other information sources provides us with architectural bounds on when information can be integrated between modalities – either lexically or sub-lexically.

The following section provides a brief overview of the implementation of the two alternative architectures for multimodal integration and the simulations of Experiment 1 and 2 to generate predictions about behaviour resulting from different patterns of information integration. From each model we are able to extract a detailed prediction of the time course of fixation towards each category of item presented within the two experiments, analysing the onset and offset of any visual, semantic and/or rhyme effects, their relative magnitudes and, by comparing across experiments, the effect of additional competition on any rhyme effect observed. This is then followed by a description of the two experimental visual world studies. Results of the simulations are then evaluated in light of experimental findings and their consequences for language mediated eye gaze research and, more broadly, the multimodal architecture supporting spoken word processing are discussed.

In brief, we observe that when visual and semantic competitors are presented alongside phonological rhyme competitors, rhyme effects are no longer observed. Such data proves more consistent with predictions generated by a fully interactive architecture, represented in this study by the MIM, which predicts small rhyme effects that are then reduced when visual and semantic competitors are presented simultaneously. However, a system in which multimodal information integration is restricted to the lexical level, represented in this study by the extended TRACE model, by contrast consistently predicts larger rhyme effects that increase in the presence of visual and semantic competitors. Thus, our data supports the position that information is able to interact across modalities sublexically during language processing.

Simulating the effects of multimodal competition on phonological rhyme overlap in a fully interactive model of language mediated visual attention

The Multimodal Integration Model (MIM) of language mediated visual attention

The Multimodal Integration Model (Smith et al., 2013; 2014a; 2014b) of language mediated visual attention was used for simulations within this study. Previous studies have demonstrated the model's ability to capture a broad range of word level properties of language mediated visual attention (see Table 1). The architecture, representations and training procedure replicated those described in Smith et al. (2013)[1]. An overview of the implementation is provided below, for a full description of the motivation for and structure of the model, refer to Smith et al. (2013).

Architecture

The MIM utilises the parallel distributed processing framework (see Rogers & McClelland, 2014; Rumelhart, McClelland, & the PDP Research Group, 1986). The network consists of layers of processing units connected via weighted connections. The architecture of the model is displayed in Fig. 1. A layer of 80 units defines the visual layer. This layer provides input of visual information to the network from four locations (each represented by 20 units) in the visual field. A layer of 60 units provides input of phonological information to the network. This layer is divided into six phoneme slots, with each slot consisting of 10 units each sensitive to a specific phonological property of an utterance at a specific temporal location. Units in both phonological and visual layers are fully connected in a forward direction to a central integrative layer. The integrative layer consisted of 400 units and is fully self-connected. The integrative layer is also fully connected to both a semantic layer and an eye layer, in both forward and backward directions. The semantic layer consists of 200 units each of which are sensitive to a specific semantic property. The eye layer consists of four units with each unit encoding the probability that the model directs gaze to one of the four locations in the visual field.

Representations

24 artificial corpora were constructed, with each used to train and test a single simulation run of the model, this ensured that relationships within and between modalities were controlled. Each corpus consisted of 200 words, with each word assigned a unique phonological, semantic and visual representation. All words within the corpus were six phonemes in length. A phoneme inventory consisting of 20 phonemes was constructed, with each phoneme represented by a unique 10 unit binary phonological feature vector. Each phonological feature was assigned with p (active) = .5. Phonological representations were constructed by pseudo randomly sampling a unique sequence of six phonemes from the phoneme inventory to create each word. Controls ensured no more than 2 consecutive phonemes were shared between words (other than in the case of phonological rhyme competitors, see Table 2). Visual representations were unique 20 unit binary feature vectors, with each unit representing the presence or absence of a specific visual feature. Visual features were assigned with p(active) = .5. Semantic representations by contrast were sparsely distributed, with each word pseudo-randomly assigned a unique set of eight semantic features from a possible 200. A maximum of 1 semantic property was shared between items (other than in the case of semantic competitors where 4 properties were shared, see Table 2).

Constructing artificial corpora ensured that we controlled the relations between the stimuli, specifically the relationships between competitors, targets and unrelated items. Embedded within the corpus were 20 sets of items that shared increased overlap in either semantic, visual or phonological dimensions. Each of the 20 sets contained a target, a phonological competitor, a semantic competitor and a visual competitor. Each 'target' word shared the final three of its six phonemes with the phonological rhyme competitor. As Table 2 indicates, overlap between phonologically unrelated items was small, with a mean 0.31 phonemes overlapping between words. Similarly, in natural language vocabularies, overlap is small, for the 9374 words of length 6 phonemes in English from the CELEX database, the mean overlap in phonemes was .39 between any pair of words. Semantic competitors were defined by sharing four of their eight semantic features with the target and visual competitors a minimum of 5 of their 10 visual features with the target. This ensured that in all dimensions the distance between competitor and target was half that between competitor and an unrelated item (see Table 2).

Training

In training the model we assume that individuals learn associations between representations of an item across modalities through repeated, simultaneous exposure to multiple representational forms of an item. Networks were trained on four cross modal mapping tasks: object recognition; spoken word comprehension; speech motivated orientation; and semantically motivated orientation. Time in the model was represented by the flow of information across weighted connections between units in the network. Each training task ran for 14 time steps (ts).

For object recognition tasks, four items were randomly selected from the training corpus and their visual representations presented to the four visual input slots within the visual layer (ts = 0). One of the four items was then randomly selected as a target and the eye gaze layer unit corresponding to the location of the target's visual representation in the visual layer was fully activated (ts = 0). Visual input and eye gaze layer activation remained fixed across the training trial while random time invariant noise was provided as an input to the phonological layer. At time step 3 until the end of the trial (ts = 14) the semantic representation of the target was presented to the semantic layer and error back propagated.

Spoken word comprehension tasks involved randomly selecting an item from the corpus as a target. The phono-

[1] Model used within this study replicates the 'noisy learning environment' implementation described within Smith et al. (2013).

Table 1
Table presents data recorded in the Visual World Paradigm that the MIM has previously been demonstrated to capture (Smith et al., 2013, 2014b).

Study	Scene			
Authors (year)	Item 1	Item 2	Item 3	Item 4
Allopenna et al. (1998)	Target	**Onset**	**Rhyme**	Distractor
Dahan and Tanenhaus (2005)	Target	**Visual**	Distractor	Distractor
Huettig and Altmann (2007)	**Visual**	Distractor	Distractor	Distractor
Yee and Sedivy (2006)	Target	**Semantic**	Distractor	Distractor
Huettig and Altmann (2005)	**Semantic**	Distractor	Distractor	Distractor
Mirman and Magnuson (2009)[a]	Target	**Near Semantic**	**Far Semantic**	Distractor
Huettig and McQueen (2007)[b]	**Onset**	**Semantic**	**Visual**	Distractor

Notes: Item 1–4 indicate the relationship of each of the four objects presented in the visual display of each study to the spoken target word. Observed competitor effects are indicated in bold type. Onset = phonological onset competitor; Rhyme = phonological rhyme competitor; Visual = visual competitor; Semantic = semantic competitor.

[a] Near and Far semantic competitors presented on separate trials.

[b] Experiment 1.

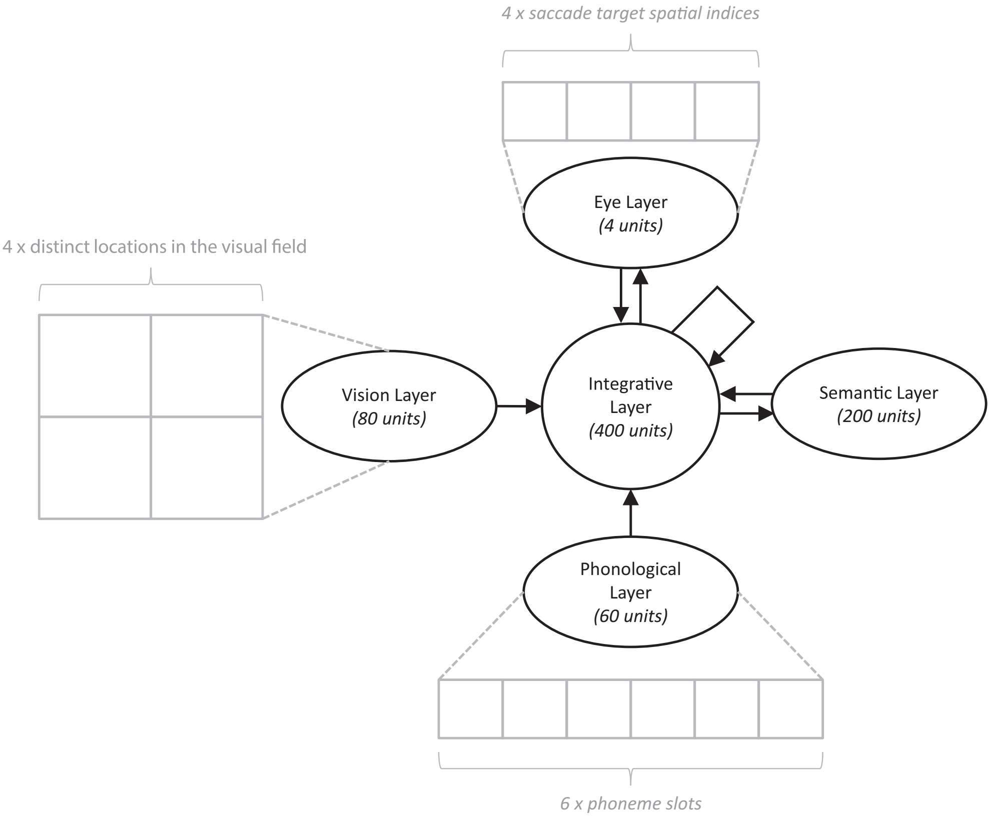

Fig. 1. Architecture of the multimodal integration model of language mediated visual attention.

logical representation of the target item was then over time (from ts = 0) presented to the phonological layer of the network, with an additional phoneme presented at each subsequent time step. To simulate exposure to noise in the auditory input within the learning environment the binary value of each unit within the phonological representation of the target was switched (i.e. 0 -> 1 or 1 -> 0) with $p = .2$ (see Smith et al., 2013). Random time invariant noise was presented as input to the visual layer during such trials, while no constraints were placed on eye layer activity. At time step 5 the semantic representation of the target was presented to the semantic layer and error backpropagated until the end of the training trial (ts = 14).

For phonological orientation tasks, four items were randomly selected and their visual representations presented as input to the visual layer (ts 0–14). One of the four items was randomly selected as a target. The target's

Table 2
Details of relationships between targets, competitors and unrelated distractors embedded within artificial corpora.

Representation	Item type	Constraint (features shared with target)	Cosine distance
Phonological	Competitor	Final 3 of 6 phonemes	0.259
	Unrelated	Max. 2 consecutive phonemes	0.496
Semantic	Competitor	4 of 8 semantic features	0.500
	Unrelated	Max. 1 semantic feature	0.959
Visual	Competitor	Min. 5 of 10 visual features	0.264
	Unrelated	Features shared with $p = (.5)$	0.506

phonological representation was then presented over time (from ts = 0) as input to the phonological layer, with an additional phoneme presented at each subsequent time step. As in word comprehension tasks, to simulate exposure to noisy auditory signals in the learning environment the value of each unit in the target's phonological representation was switched with $p = .2$. No constraints were placed on activity in the semantic layer. At time step 5 (point of phonological disambiguation) the eye layer unit corresponding to the location of the target's visual representation was required to be fully activated and error backpropagated until the end of the training trial (ts = 14).

Finally, semantic orientation trials followed a similar procedure. Again four items were randomly selected from the corpus and their visual representations presented as input to the visual layer (ts 0–14). One of these four items was randomly selected as a target and its semantic representation presented to the semantic layer (ts 0–14). Random time invariant noise was presented to the phonological layer throughout this trial. At time step 2 the eye layer unit that corresponded to the location of the visual representation of the target was required to be fully activated and error backpropagated until the end of the training trial (ts = 14).

We assume that speech motivated orientation is less frequent in the learning environment than object recognition, spoken word comprehension and semantically motivated orientation and therefore this task was four times less likely to occur during training. Given this constraint training tasks were randomly interleaved.

Simulations were conducted using Mikenet version 8.0 developed by M.W. Harm (www.cnbc.cmu.edu/~mharm/research/tools/mikenet/), a collection of libraries written in the C programming language for implementing and training connectionist networks. Connection weights within the model were initialised with random weights from the uniform distribution [−0.1, 0.1]. Recurrent backpropagation (learning rate = 0.05) was used during training to adjust weights within the network using the continuous recurrent backpropagation through time training algorithm provided in Mikenet (crbp.c) which implements Pearlmutter (1989). Unit activation was calculated using a logistic activation function and sum squared error was used to calculate error. Time within the network was modelled using an integration constant of 0.25 with 14 samples during training and 30 samples during test simulations of visual world conditions (time steps of test trials are reported relative to word onset [i.e. word onset = ts 0]). Additional time was provided during test simulations to allow insight into the time course of interaction of information between modalities in the model. All other parameters were set to the default values implemented in Mikenet version 8.0. A total of 1,250,000 training trials were performed before the model was exposed to test conditions. Once trained all networks performed spoken word comprehension and object recognition tasks accurately (i.e. semantic layer activity was closest in terms of cosine distance to that of the target) for all items in the training corpus. On orientation tasks the model looked to the location of the target on at least 3 of 4 test trials for 99.75% (speech motivated orientation) and 100% (semantically motivated orientation) of items. 24 simulation runs of the model were performed, each initiated with a different initial random seed. Sections 'Simulation 1: Simulating effects of phonological rhyme overlap' and 'Simulation 2: Simulating effects of multimodal competition' report mean behaviour calculated across all 24 simulation runs of the model.

Simulation 1: Simulating effects of phonological rhyme overlap in the MIM

Previous visual world studies demonstrate that phonological rhyme overlap exerts an influence on language mediated visual attention under conditions in which phonology provides the only dimension in which auditory and visually presented stimuli are related. We first examine the model's sensitivity to phonological rhyme overlap when presented with scenes containing a single rhyme competitor and three unrelated items.

Procedure

Test trials lasted a total of 30 time steps (ts −5 to 24). The visual representations of four objects were presented to the visual layer at trial onset (ts = −5) and remained present until the end of the trial (ts = 24). Three of the items were unrelated to the upcoming target word, i.e., controlled low level of overlap with the target in visual, semantic or phonological dimensions (see Table 2). The fourth item was a phonological rhyme competitor in that it shared the final three phonemes of its phonological representation with the upcoming target word. The network was then free to cycle for five time steps (ts −5 to −1) to allow pre-processing of the visual information (replicating previous visual world studies in which a preview of the visual display is often provided, see Huettig & McQueen, 2007). At time step 0 the phonological representation of the target word began to unfold with an additional phoneme presented at each subsequent time step to the phonological layer. Unlike in training, no noise was applied to the phonological input of the target representation. Activation in the eye layer was recorded throughout the trial. The location in the visual field fixated by the model at a given time point was recorded as the location associated with the most activated unit in the eye layer at the given point in time. This procedure was followed for all rhyme competitor and target pairs within the corpus ($n = 20$) with rhyme competitors and distractors tested in all possible combinations of location ($n = 24$) resulting in a total of 480 test trials per simulation run of the model.

Results

Fig. 2 presents the change from word onset (ts = 0) in the probability of fixating rhyme competitors and unrelated distractors. To allow us to compare the same items across conditions (i.e. difference in probability of fixating a rhyme competitor compared to an unrelated distractor in the presence or absence of visual and semantic competitors) we randomly selected one of the three unrelated distractors from each display and report the probability of fixating these items as the probability of fixating an unrelated distractor.

To examine whether looks to phonological rhyme competitors exceeded looks to unrelated distractors, for analysis we divided the 30 time step test trial into five equal time windows. We then compared fixation behaviour displayed by the model in the baseline time window (ts −5 to 0), the 6 time steps from trial onset to word onset, to fixation behaviour displayed by the model in each of the four time windows post word onset (ts 1–6; ts 7–12; ts 13–18; ts 19–24). For each window we calculated the empirical log odds of fixating each category of object within the display (i.e., rhyme competitor, unrelated distractor). This measure avoids issues arising from calculating estimates based on proportional data (see Jaeger, 2008). Our dependent measure was the difference between the log-odds of fixating the phonological rhyme competitor and the log-odds of fixating the unrelated distractor. This reflects the difference in fixation of competitor objects as a consequence of representational overlap. We used linear mixed effect models to examine whether gaze differed as a consequence of phonological rhyme overlap in the time windows post word onset relative to levels of fixation in the baseline time window. Mixed effects model analysis was performed using the R (version 3.1.0; R Development Core Team, 2009) libraries lme4 (version 1.1-6; Bates, Maechler, Bolker, & Walker, 2015). The model constructed applied the maximal random effect structure (Barr, Levy, Scheepers, & Tily, 2013), the fixed effect time window and random effects of model simulation run ($n = 24$) and item ($n = 20$), including random intercepts and slopes for time window both by simulation run and item. To derive p-values we assumed t-values were drawn from a normal distribution (Barr, 2008).

Examining parameter estimates within the model revealed that in the first time block that followed word onset (ts 1–6) phonological rhyme competitors were fixated marginally less than unrelated distractors relative to the baseline time window ($\beta = -0.082$, $t = -1.72$, $p = .086$). In the second time window (ts 7–12), this trend reversed with rhyme competitors fixated more than unrelated items ($\beta = 0.385$, $t = 3.02$, $p = .003$). This increased rhyme effect remained in the final two time windows ts 13–18 ($\beta = 0.486$, $t = 3.68$, $p < .001$) and ts 19–24 ($\beta = 0.480$, $t = 3.53$, $p < .001$).

Summary

Results of Simulation 1 demonstrate that the MIM displays sensitivity to phonological rhyme overlap when presented with scenes containing a single rhyme competitor amongst unrelated items with effects predicted to emerge post word offset. This replicates previous behavioural findings that language mediated eye gaze is sensitive to phonological rhyme overlap between spoken target words and visually displayed objects (Allopenna et al., 1998; McQueen & Huettig, 2012; McQueen & Viebahn, 2007). Further, the model demonstrates that a fully interactive alignment model of spoken word processing (MIM) is able to generate phonological rhyme effects (Magnuson et al., 2000, 2003; Smith et al., 2013; cf. Allopenna et al., 1998).

Simulation 2: Simulating effects of multimodal competition in the MIM

A second set of simulations examined the relative influence and timing of effects of phonological rhyme, semantic and visual overlap on eye gaze within the MIM and the effect of additional competition from visual and semantic competitors on phonological rhyme effects that were exhibited in Simulation 1.

Procedure

Simulation 2 followed the same training and testing procedure as outlined for Simulation 1 (see Section 'Simulation 1: Simulating effects of phonological rhyme overlap in the MIM – Procedure'), however test scenes now contained a rhyme competitor, a semantic competitor, a visual competitor and an unrelated distractor (scenes contained the same rhyme competitor and unrelated distractor pairs as analysed in Section 'Simulation 1: Simulating effects of phonological rhyme overlap in the MIM – Results'). Again simulations were run for all target and competitor sets embedded within the corpus ($n = 20$) with sets tested in all possible combinations of location ($n = 24$) resulting in a total of 480 test trials per simulation run. Results report the probability of fixating an item at any given time point, this is taken as the proportion of trials on which at that given point in the trial the eye layer unit associated with location of the given object is the most activated unit in the eye layer.

Results

The change in the probability of fixating each category of item (i.e., rhyme competitor, semantic competitor, visual competitor and unrelated distractor) from word onset is presented in Fig. 3. Visual inspection suggests a rapid increase in fixation of visual competitors shortly after word onset, with increased fixation of semantic competitors emerging slightly later and at lower levels. Fixation of phonological rhyme competitors also appears to depart from unrelated distractor levels however this appears later than semantic and visual competitors and is a weaker effect.

We used the same procedure for analysis of Simulation 2 as used in Simulation 1. However, scenes in Simulation 2 contained three competitors rather than a single competitor in Simulation 1. We therefore compared separately for

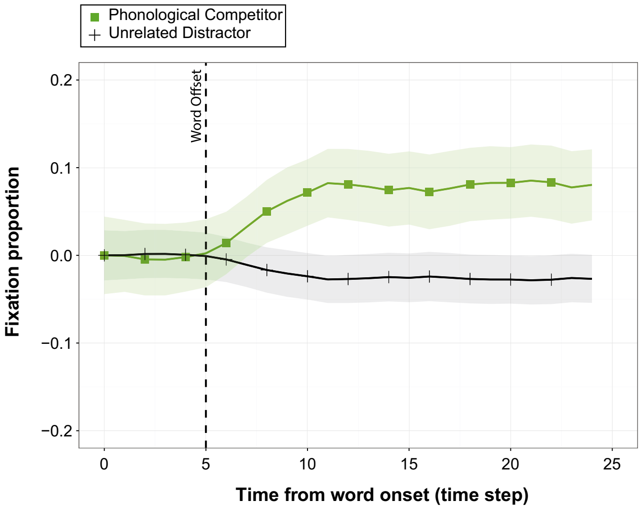

Fig. 2. Change in proportion of fixations from word onset (ts = 0) displayed by the multimodal integration model to items in visual displays containing a rhyme competitor and three unrelated distractors. Shaded areas define 95% confidence intervals.

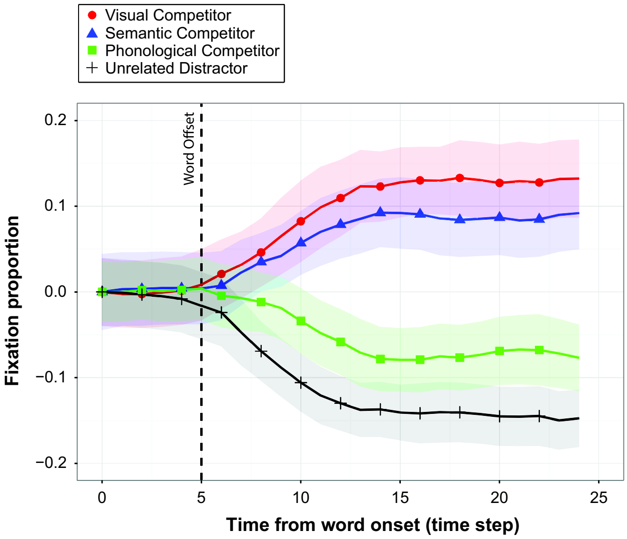

Fig. 3. Change in the proportion of fixations from word onset displayed by the multimodal integration model to items in visual displays containing a rhyme competitor, a visual competitor, a semantic competitor and an unrelated distractor. Shaded areas define 95% confidence intervals.

each category of competitor (visual, semantic, rhyme) the difference in empirical log odds of fixating the given competitor and the unrelated distractor in each time window post word onset (ts 1–6, ts 7–12, ts 13–18, ts 19–24) to the difference observed in the baseline time window (ts −5 to 0), the 6 time steps from trial onset to word onset. For analysis we used linear mixed effect models with a fixed effect of time window and random effects of model simulation run ($n = 24$) and item ($n = 20$), including random intercepts and slopes for time window both by simulation run and item.

This analysis revealed that phonological rhyme, visual, and semantic competitors were fixated above unrelated distractor levels in windows ts 7–12 (rhyme: $\beta = 0.284$, $t = 2.90$, $p = .004$; visual: $\beta = 0.727$, $t = 8.27$, $p < .001$; semantic: $\beta = 0.666$, $t = 6.99$, $p < .001$), ts 13–18 (rhyme: $\beta = 0.289$, $t = 2.90$, $p = .004$; visual: $\beta = 1.265$, $t = 12.2$, $p < .001$; semantic: $\beta = 1.107$, $t = 9.41$, $p < .001$) and ts 19–24 (rhyme: $\beta = 0.345$, $t = 3.60$, $p < .001$; visual: $\beta = 1.311$, $t = 14.80$, $p < .001$; semantic: $\beta = 1.138$, $t = 9.01$, $p < .001$). While in the first time block post word onset (ts 1–6) there was no difference between competitors and unrelated distractors (rhyme: $\beta = 0.014$, $t = 0.244$, $p = .807$; visual: $\beta = -0.045$, $t = -0.864$, $p = .388$; semantic: $\beta = 0.003$, $t = 0.052$, $p = .959$).

To examine whether the magnitude of competitor effects differed we used the same analysis technique to test in each of the post word onset time windows (ts 1–6, ts 7–12, ts 13–18, ts 19–24) relative to the baseline time window (6 time steps prior to word onset, ts −5 to 0) whether: the log odds of fixating the visual competitor differed from the log odds of fixating the semantic competitor; the log odds of fixating the visual competitor differed from the log odds of fixating the rhyme competitor; and the log odds of fixating the semantic competitor differed from the log odds of fixating the rhyme competitor. This analysis did not reveal a significant difference between fixation of the semantic competitor compared to fixation of visual competitors in any time window relative to the baseline window (ts 1–6: $\beta = -0.048$, $t = -0.765$, $p = .444$; ts 7–12: $\beta = 0.061$, $t = 0.585$, $p = .558$; ts 13–18: $\beta = 0.159$, $t = 1.251$, $p = .211$; ts 19–24: $\beta = 0.173$, $t = 1.33$, $p = .185$). Visual and Semantic competitors were however fixated more than rhyme competitors relative to baseline levels in all time windows post word onset other than ts 1–6 (Semantic vs. Rhyme: ts 1–6: $\beta = -0.011$, $t = -0.204$, $p = .838$; ts 7–12: $\beta = 0.383$, $t = 3.92$, $p < .001$; ts 13–18: $\beta = 0.818$, $t = 7.39$, $p < .001$; ts 19–24: $\beta = 0.793$, $t = 7.52$, $p < .001$; Visual vs. Rhyme: ts 1–6: $\beta = -0.059$, $t = -1.16$, $p = .248$; ts 7–12: $\beta = 0.443$, $t = 3.66$, $p < .001$; ts 13–18: $\beta = 0.976$, $t = 7.55$, $p < .001$; ts 19–24: $\beta = 0.966$, $t = 7.74$, $p < .001$).

Finally, using mixed effects models we analysed whether the difference in empirical log odds of fixating the phonological rhyme competitor and empirical log odds of fixating the unrelated distractor differed between Simulation 1 and Simulation 2. Did the presence of visual or semantic competitors influence the magnitude of the phonological rhyme effect? This was performed using a model with fixed effects of time window and scene (Scene 1: rhyme competitor and unrelated distractors only; Scene 2: rhyme competitor, semantic competitor, visual competitor and unrelated distractor) and random effects of model simulation run ($n = 24$) and item ($n = 20$), including random intercepts and slopes for both time window and scene both by simulation run and item. Analysing the rhyme effect in the baseline time window (6 time steps prior to word onset, ts −5 to 0) in relation to that observed in ts 1–6 revealed no main effect of time window ($\beta = -0.034$, $t = -0.69$, $p = .490$), although there was a marginal main effect of scene ($\beta = 0.107$, $t = 1.67$, $p = .096$) and a marginal interaction between time window and scene ($\beta = 0.096$, $t = 1.89$, $p = .059$) suggesting that the presence of visual and semantic competitors increased the rhyme effect marginally in this early window. By contrast comparing the rhyme effect observed in the baseline time window to that observed in later time windows ts 7–12, ts 13–18 and ts 19–24 revealed for all later windows a main effect of time window (ts 7–12: $\beta = 0.334$, $t = 3.15$, $p = .002$; ts 13–18: $\beta = 0.388$, $t = 3.62$, $p < .001$; ts 19–24: $\beta = 0.413$, $t = 3.82$, $p < .001$), no main effect of scene (ts 7–12: $\beta = 0.008$, $t = 0.049$, $p = .868$; ts 13–18: $\beta = -0.039$, $t = -1.16$, $p = .248$; ts 19–24: $\beta = -0.009$, $t = -0.244$, $p = .807$) and a significant negative interaction term between time window and scene (ts 7–12: $\beta = -0.102$, $t = -2.27$, $p = .023$; ts 13–18: $\beta = -0.197$, $t = -4.32$, $p < .001$; ts 19–24: $\beta = -0.135$, $t = -2.91$, $p = .004$) demonstrating that the presence of visual and semantic competitors reduced the rhyme effect in later time windows.

Summary

In summary the MIM predicts that all items that overlap with the spoken target word in terms of visual properties, semantic properties or phonological rhyme will be fixated above unrelated distractor levels. All effects emerged gradually post word offset and within the same time window post word offset (ts 7–12). With regard to behavioural data, the MIM with interactivity at all stages of processing, predicts that visual and semantic effects will be of a similar magnitude, although beta estimates were numerically higher in all post word offset time windows for visual competitors. In relation to rhyme competitor effects both visual and semantic competitor effects were predicted to be far greater, with differences emerging shortly after word onset and increasing across the remainder of the test window. Comparisons of rhyme effects observed in the presence (Simulation 2) and absence (Simulation 1) of visual competitors generate the prediction that the presence of visual and semantic competitors will weaken the effect of phonological rhyme overlap on fixation behaviour in time windows post word offset.

Simulating the effects of multimodal competition on phonological rhyme overlap in a lexical level cascading model of language mediated visual attention

In this section we detail the predictions generated by a cascaded architecture in which activation from visual and sematic levels connects with phonological activation at the lexical level as simulated by the word level nodes of the TRACE (McClelland & Elman, 1986; Spivey, 2007) model. The architecture of our implementation of this extended TRACE model is presented in Fig. 4. This model contrasts with the MIM in terms of when information between modalities is permitted to interact. The predictions of each model are then tested using the behavioural data of Experiments 1 and 2, below.

Implementing an influence of cascading multimodal information into the TRACE model of spoken word recognition

Simulations were performed using jTRACE (Strauss, Harris, & Magnuson, 2007). All parameters were set to default values other than the following subset that were manipulated to simulate cascading activation from visual and semantic levels entering the TRACE hierarchy to be integrated with phonological information at the lexical level. The resting level of nodes at the lexical level of the TRACE hierarchy can be affected in jTRACE by manipulating the frequency resting state (see Dahan, Magnuson, & Tanenhaus, 2001) and priming resting state parameters. These parameters alter lexical level node activation as determined by Eq. (1).

$$r_i = R + s_f[\log_{10}(c + f_i)] + s_p[\log_{10}(c + p_i)] \tag{1}$$

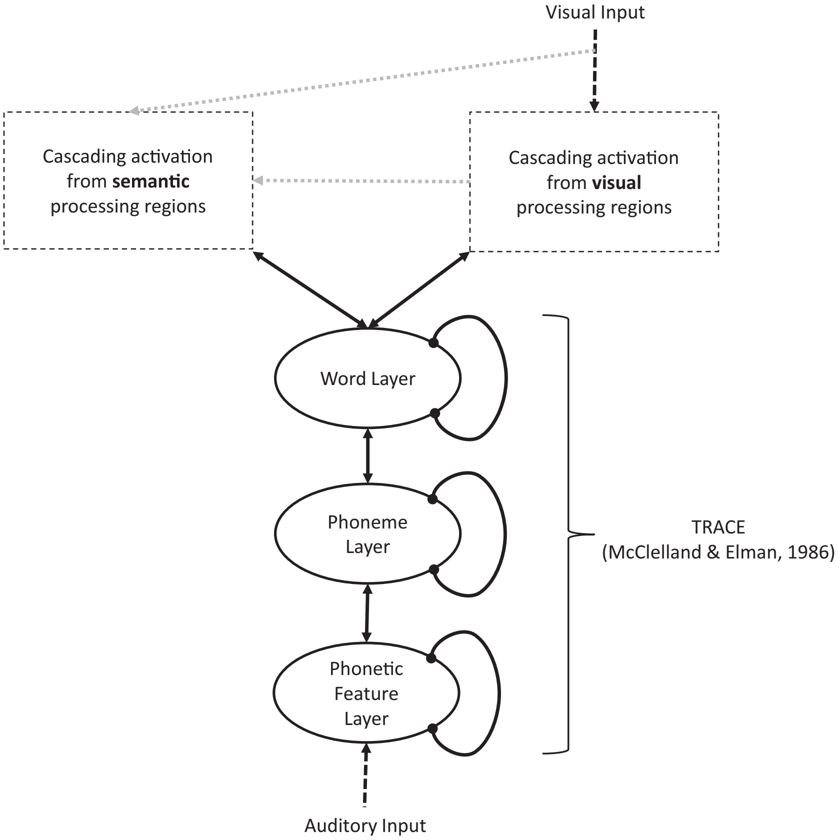
Visual Input
Cascading activation from **semantic** processing regions
Cascading activation from **visual** processing regions
Word Layer
Phoneme Layer
Phonetic Feature Layer
TRACE (McClelland & Elman, 1986)
Auditory Input

Fig. 4. Architecture of the extended multimodal TRACE model.

In Eq. (1) r_i is the resting activation for unit i, R is the default resting level for all units, s_f is the frequency resting level scaling constant, c is a constant that ensures the value within parenthesis is greater than 0, s_p is the priming resting level scaling constant, f_i is the frequency of item i, p_i is the priming value for item i. In our simulations f_i and p_i were used to represent the relative level of representational overlap at visual and semantic levels between a given item and the target word, and thus represents the relative magnitude of activation cascading from visual and semantic levels to activate the node representing item i at the lexical level. A positive linear function applied to s_f and s_p ensured that cascading activation from visual and semantic levels ramped up over time as activation of associated representations in these modalities increased. We implemented the assumption that time is required for activation to cascade between modalities. This was done by ensuring that activation generated at visual and semantic levels by overlap in cascading activation from visual and auditory signals began influencing activation at the lexical level within TRACE six time steps after the onset of the spoken target word (equal to the time taken for a single phoneme to unfold).

Pilot simulations explored the parameter space in order to identify a range of values for f_i, p_i, s_f and s_p able to generate behaviour consistent with that recorded in the data sets detailed in Table 1, which a fully interactive parallel processing system (MIM: Smith et al., 2013, 2014b) has previously demonstrated an ability to replicate (see Appendix B for further details of pilot simulations conducted with the extended TRACE model that demonstrate the model's ability to generate visual and semantic competitor effects in both the presence and absence of the target item in addition to the complex time course of fixation behaviour generated by multi-competitor scenes as observed in Allopenna et al. (1998) [phonological onset competitor, phonological rhyme competitor & target] and Huettig and McQueen (2007) [phonological onset competitor, visual competitor & semantic competitor]).

In our initial parameterisation of TRACE we assume that the resting state of nodes at the lexical level corresponding to items in the visual display is equal at word onset. As the spoken target word unfolds this increases activation of phonologically related words. Activation of such words at the lexical level then cascades to activate the visual and semantic properties of these items. At the same time, information relating to the visually displayed items is also cascading from visual levels to constrain activation of the visual and semantic properties. Thus at the lexical level, post target word onset, cascading activation from visual and semantic levels should increase activation of items that share properties at the visual and semantic level that are supported both by the incoming visual and auditory signal.

In our simulations we therefore aimed to model the change from word onset in the nature of activation cascading from visual and semantic levels to affect lexical level activation. This is performed by manipulating parameters f_i, p_i, s_f and s_p in jTRACE. In these simulations we use f_i to define the magnitude of the increase, relative to word onset, of activation cascading from visual levels to activate the lexical level node corresponding to item *i*. p_i is used to define the magnitude of the increase, relative to word onset, of activation cascading from semantic levels to activate the lexical level node corresponding to item *i*. s_f is a scaling factor defining the level of influence of cascading activation from visual levels on activation of all lexical level nodes. s_p is a scaling factor defining the level of influence of cascading activation from semantic levels on all lexical level nodes. We determined the relative values of parameters f_i, p_i for each category of item (target, visual competitor, semantic competitor, phonological onset competitor, phonological rhyme competitor, unrelated distractor) by first assigning values to the items that overlapped maximally and minimally with the shared auditory and visual signals, the target item and the unrelated distractor.

In target present scenes the lexical level node of the target item is supported maximally by information cascading from phonological processing levels, given that its complete phonological form is present in the auditory input signal. Similarly, cascading activation from visual and semantic levels to lexical level nodes can also be assumed to increase maximally post word onset as the target's visual and semantic properties are likely fully activated by activation cascading both top down from lexical levels and bottom up from visual levels given the presence of the target's full visual form in the visual input. We therefore assigned the ceiling value of 1000 for parameters f_i and p_i (see Dahan et al., 2001) for target items when present in the display. Conversely, although the visual input signal contains the visual form of the unrelated distractor, there is no additional support for this item in the auditory signal. Thus, from word onset, there should be no increase in activation of visual or semantic properties of the unrelated distractor. For this reason parameters f_i and p_i were assigned a value of 0 for unrelated distractors. All items in the corpus that were not present in the visual display were also assigned a value of 0 for both parameters f_i and p_i.

By contrast competitor items each receive additional support from cascading activation initiated post word onset. In the case of the visual competitor, the auditory signal increases activation of the target's visual properties, which are shared with the visual competitor, thus we assume this also increases the amount of activation cascading from visual levels to activate the visual competitor at the lexical level. We therefore assigned visual competitors an f_i value of 500 ($p_i = 0$), half that of the target item. Similarly, for semantic competitors, activation of the semantic properties corresponding to the target word, which are also shared with the semantic competitor, are assumed to increase activation cascading from semantic levels to activate the semantic competitor at the lexical level. Therefore, p_i was assigned a value of 500 for all semantic competitors ($f_i = 0$), half that of the target. These ratios were motivated by the results of behavioural rating studies (see Table 5) and Huettig and McQueen (2007, Table 2) which showed that the ratio of the similarity between unrelated distractor and target compared to visual or semantic competitor and target was approximately 0.5 (Huettig & McQueen, 2007: Visual/Unrelated = 0.51, Semantic/Unrelated = 0.50; Smith, Monaghan & Huettig, current study: Visual/Unrelated = 0.41, Semantic/Unrelated = 0.47).

Given such an architecture, we can also assume that cascading activation from visual and semantic levels also increases post word onset, relative to unrelated distractor levels, for both phonological onset and phonological rhyme competitors should they be present in the visual display. As discussed in Section 'Models of multimodal integration during speech processing', previous applications of TRACE to model visual world data (e.g. Allopenna et al., 1998) demonstrated that items that share phonological properties in their rhyme or onset with the target word are activated above unrelated items at the lexical level. Within this architectural framework we therefore assume that this activation also then cascades to activate the visual and semantic properties of such phonologically related items, properties that are also supported by cascaded activation from the visual input given the presence of their visual form in the visual display. Further, these previous simulations also demonstrate that the location (onset or rhyme) of the phonological overlap will determine the magnitude and onset of this cascading activation.

Allopenna et al. (1998) shows that activation of lexical level nodes for phonological onset competitors increases at a rate identical to targets, prior to the speech signal disambiguating between the two items. Following disambiguation, activation of the onset competitor at the lexical level decreases rapidly. Therefore, to simulate such conditions in the extended TRACE model, for scenes in which onset competitors are present, in the period prior to phonological disambiguation the onset competitors parameters are f_i and $p_i = 1000$. In the period post disambiguation parameters f_i and $p_i = 0$. By contrast activation of the lexical level node corresponding to the rhyme competitor only increases above unrelated distractor levels once phonological information carried in the rhyme becomes available. Further the overall level of activation reached by the rhyme competitor is lower than that obtained by the onset competitor as earlier activation of onset competitors inhibits later activation of the rhyme competitor. However, as there is activation of the rhyme competitor at the lexical level above unrelated distractor levels, we can assume this additional activation cascades to activate semantic and visual properties associated with the rhyme competitor. If the rhyme competitor is present in the visual scene, this cascading activation to visual and semantic levels will be supported by cascading activation initiated by the presence of the rhyme competitor's visual form in the visual signal. We therefore assign parameters f_i and $p_i = 50$ for rhyme competitors so as to simulate the small levels of activation cascading from visual and semantic levels to influence activation at the lexical level. We believe a value of 5% of the target's activation is a conservative estimate given that in Allopenna et al. (1998) the lexical level nodes corresponding to rhyme competitors are at their peak activated to approximately 10% the level of the target item. Note that increasing this value of f_i and p_i for rhyme competitors would have the effect of further increasing the activation of the phonological competitor.

By applying a linear function to scaling factors s_f and s_p we controlled the onset and level of activation cascading over time from visual and semantic levels, respectively. For all items other than rhyme competitors activation of their lexical level nodes due to cascading activation from visual levels increased from 6 time steps post word onset as scaling factors s_f increased linearly from 0 to 0.2 over the course of 24 time steps (number of time steps required for full phonological form of target word to unfold). Activation from semantic levels increased in an identical manner but with the onset delayed by 6 time steps in order to simulate a delayed effect of semantic activation (see Huettig & McQueen, 2007). For rhyme competitors activation entering the lexical level from both visual and semantic levels increased 12 time steps post word onset, with scaling factors s_f and s_p increased linearly from 0 to 0.2 over the course of 24 time steps. Thus, the onset of such increased activation occurs six time steps after the onset of the first phoneme that overlaps with the target word.

A second parameterisation of the extended TRACE system was also tested. This second parameterisation aimed to maximise the likelihood of observing effects of visual and semantic competition on the rhyme effect. In this parameterisation, activation cascading from visual levels to influence lexical nodes occurred at the same time step as that cascading from semantic levels. In the case of rhyme competitors the onset of activation occurred 6 time steps later than all other items. The scaling constant determining the influence of semantic and visual activation (s_f and s_p) was also increased linearly as described above yet to an increased level of 0.5 (pilot studies explored values for s_f and s_p beyond this level yet this generated an increasingly worse fit to the data sets described in Table 1). In the second parameterisation of the system to maximise competition all related items (target, onset competitor, rhyme competitor, semantic competitor, visual competitor) received feedback from visual and semantic levels, even though they may not be present in the visual display.

As in previous studies that have applied TRACE to model visual world data, the Luce choice rule (Luce, 1959) was applied to raw lexical node activations for each item present in the visual display, the result of which was taken to represent the probability of fixating each displayed item.

We used the default corpus provided with jTRACE which was supplemented with additional words to create 10 distinct stimuli sets (see Appendix Table B1). Each set included a target word, a phonological onset competitor, a phonological rhyme competitor, a visual competitor, a semantic competitor and an unrelated distractor. All words were four phonemes in length. Rhyme competitors shared all but their initial phoneme with the target. Onset competitors shared their initial two phonemes with the target. Controls (see Table 3) ensured that rhyme competitors, visual competitors and semantic competitors did not differ from unrelated distractors in their cohort density ($t < 1.01$, $p > .32$), while visual and semantic competitors did not differ from unrelated distractors in the number of phonemes shared with the target (Table 3, shared phonemes: $t < 0.69$, $p > .49$). Further visual competitors, semantic competitors and unrelated distractors did not have any shared phonemes in the same location as the target (see Table 3, phoneme overlap).

Simulation 3: Simulating effects of phonological rhyme overlap in TRACE

We used the extended TRACE model to first generate predictions for how fixation would be distributed towards objects in a scene that contained a single rhyme competitor accompanied by only unrelated items (i.e. conditions simulated in MIM in Section 'Simulation 1: Simulating effects of phonological rhyme overlap in the MIM').

Procedure

In total, ten trials were run for each parameterisation of the TRACE system (see Table 4, Scene 1), with one trial for each of the rhyme competitor and unrelated distractor pairings defined in the ten stimuli sets (see Appendix Table B1). Trials lasted a total of 70 time steps (ts −6 to 64) allowing time for activation to cascade across levels within the network. All time steps are recorded relative to word onset (i.e. word onset = time step 0) although 6 steps elapsed prior to word onset (ts −6 to −1). It took 6 time steps for each phoneme to unfold.

Results

Fig. 5 displays the change from word onset in the probability of fixating each category of item (rhyme competitor, unrelated distractor) averaged over the 10 test trials for parameterisation 1 (Fig. 5A) and parameterisation 2 (Fig. 5B) of the extended TRACE model.

Table 3
Controls on TRACE stimuli.

	Cohorts 1		Cohorts 2		Shared phonemes		Phoneme overlap	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Rhyme Competitor	26.9	11.1	5.20	2.09	3.30	0.46	3.00	0.00
Visual Competitor	30.0	8.1	7.80	3.76	1.00	0.63	0.00	0.00
Semantic Competitor	33.2	12.0	6.20	3.76	0.80	0.40	0.00	0.00
Unrelated Distractor	30.2	6.4	6.30	2.97	0.80	0.60	0.00	0.00

Table 4
jTRACE parameterisations for simulations of word processing during exposure to visual scenes for Simulation 3 (Scene 1: rhyme competitor and unrelated distractors) and Simulation 4 (Scene 2: rhyme competitor, visual competitor, semantic competitor and unrelated distractor). Time steps recorded relative to word onset (ts = 0).

Parameter Set	Item	Cascading visual activation						Cascading semantic activation					
		Activation level (f_i)		Onset		Maximal		Activation level (p_i)		Onset		Maximal	
		Scene 1	Scene 2	s_f	ts	s	ts	Scene 1	Scene 2	s_p	ts	s	ts
P1	Target	0	0	0	–	–	–	0	0	0	–	–	–
	Onset	0	0	0	–	–	–	0	0	0	–	–	–
	Rhyme	50	50	0	12	0.2	36	50	50	0	12	0.2	36
	Visual	0	500	0	6	0.2	30	0	0	0	–	–	–
	Semantic	0	0	0	–	–	–	0	500	0	12	0.2	36
	Unrelated	0	0	0	–	–	–	0	0	0	–	–	–
P2	Target	1000	1000	0	6	0.5	30	1000	1000	0	6	0.5	30
	Onset	1000	1000	0	6	0.5	30	1000	1000	0	6	0.5	30
	Rhyme	50	50	0	12	0.5	36	50	50	0	12	0.5	36
	Visual	0	500	0	6	0.5	30	0	0	0	–	–	–
	Semantic	0	0	0	–	–	–	0	500	0	6	0.5	30
	Unrelated	0	0	0	–	–	–	0	0	0	–	–	–

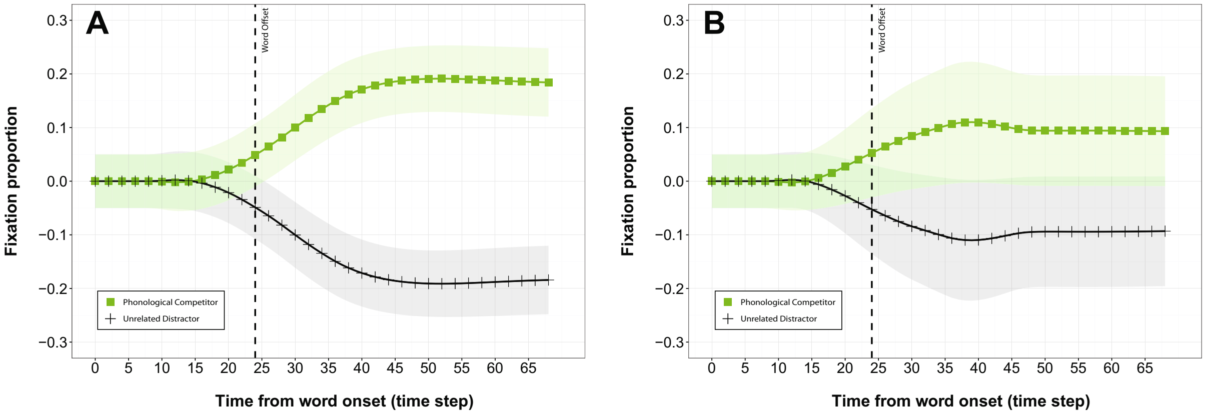

Fig. 5. Change from word onset in the probability of fixating rhyme competitors and unrelated distractors as predicted by the extended TRACE model. (A) Behaviour generated by parameterisation 1, (B) behaviour generated by parameterisation 2. Shaded areas define 95% confidence intervals.

We applied the same method of analysis as described in Section 'Simulation 1: Simulating effects of phonological rhyme overlap in the MIM' to the probabilities of fixating each category of object generated by the extended TRACE model. We first divided the 70 time step test trial into five time windows. A baseline time window was recorded as the period from trial onset to word onset (6 time steps: ts −6 to 0). The remainder of the trial, the period post word onset, was then divided into four equal length windows (ts 1–16; ts 17–32; ts 33–48; ts 49–64). For each window we calculated the empirical log odds of fixating each category of item. Our dependent measure was again the difference between the log odd of fixating the phonological rhyme competitor and the log odds of fixating the unrelated distractor. Using linear mixed effects models with a fixed effect of time window and a random effect of item ($n = 10$), including random intercepts for time window we examined whether our dependent measure differed in the baseline time window (ts −6 to 0) from that recorded in each of the time windows post word onset.

This analysis revealed that fixation behaviour did not differ from baseline levels in the time period ts 1–16 for either parameterisation of the model (parameterisation 1 (P1): $\beta = -0.001$, $t = -0.26$, $p = .792$; parameterisation 2 (P2): $\beta = 0.0001$, $t = 0.03$, $p = .976$). However, the rhyme competitor was fixated more than the unrelated distractor relative to baseline levels in time windows ts 17–32 (P1: $\beta = 0.427$, $t = 15.11$, $p < .001$; P2: $\beta = 0.408$, $t = 3.65$, $p < .001$), ts 33–58 (P1: $\beta = 1.309$, $t = 28.58$, $p < .001$; P2: $\beta = 0.795$, $t = 3.98$, $p < .001$) and ts 49–64 (P1: $\beta = 1.490$, $t = 30.2$, $p < .001$; P2: $\beta = 0.723$, $t = 4.02$, $p < .001$) for both parameterisations.

Summary

The extended TRACE system fixates items that share their phonological rhyme with a spoken target word more than items that are unrelated in visual, semantic and phonological dimensions, thus replicating behaviour observed in previous visual world studies (Allopenna et al., 1998; McQueen & Huettig, 2012; McQueen & Viebahn, 2007). Both TRACE and the MIM predicted that effects emerge only post word offset. However, interpreting beta estimates as estimates of effect size indicates that the extended TRACE model predicted effects of phonological rhyme (P1: $\beta = 1.490$; P2: $\beta = 0.723$) at levels approximately twice the magnitude or greater than those predicted by the MIM ($\beta = 0.486$) when rhyme competitors are presented alongside unrelated distractors.

Simulation 4: Simulating effects of multimodal competition in TRACE

In a second set of simulations we generate predictions using the extended TRACE system for how fixations would be distributed across scenes containing a rhyme competitor, a visual competitor, a semantic competitor and an unrelated distractor (conditions simulated in MIM in Sect ion 'Simulation 2: Simulating effects of multimodal competition in the MIM').

Procedure

Again ten trials were run with each parameterisation of the model. Each trial tested the model on a distinct set that included a rhyme competitor, visual competitor, semantic competitor and unrelated distractor (see Appendix Table B1). As in previous simulations data was recorded across trials lasting 70 time steps with six time steps elapsing (ts −6 to 0) prior to word onset at time step 0.

Results

Fig. 6 displays the change from word onset in the probability of fixating each category of object (rhyme competitor, visual competitor, semantic competitor, unrelated distractor) for each parameterisation of the model (Fig. 6A = parameterisation 1, Fig. 6B = parameterisation 2) with data averaged over the 10 test trials (In simulations using parameterisation 2 the time course of fixations to visual competitors and semantic competitors followed the same trajectory and thus overlap in Fig. 6B).

The method of analysis used to analyse MIM behaviour in Section 'Simulation 2: Simulating effects of multimodal competition in the MIM – Results' was again applied to analyse the probabilities of fixating each category of object generated by the extended TRACE system. The 70 time step test trial was split into 5 time windows with the period prior to word onset (ts −6 to 0) assigned as the baseline window. We assessed separately for each category of competitor (visual, semantic, rhyme) the difference between the empirical log odds of fixating each competitor and the empirical log odds of fixating the unrelated distractor calculated in each of the time windows post word onset (ts 1–16; ts 17–32; ts 33–48; ts 49–64), which was compared to the same measure calculated in the baseline window pre-word onset (ts −6 to 0). We again used linear mixed effect models with a fixed effect of time window and a random effect of item ($n = 10$) including a random intercept for time window.

This analysis revealed that visual competitors were fixated more than unrelated distractors relative to the baseline window in all time windows post word onset for both parameterisations of the model (ts 1–16 [P1: $\beta = 0.058$, $t = 8.73$, $p < .001$; P2: $\beta = 0.152$, $t = 29.8$, $p < .001$]; ts 17–32 [P1: $\beta = 0.535$, $t = 30.43$, $p < .001$; P2: $\beta = 0.970$, $t = 93.8$, $p < .001$]; ts 33–48 [P1: $\beta = 0.687$, $t = 49.83$, $p < .001$; P2: $\beta = 0.972$, $t = 166.5$, $p < .001$]; ts 49–64 [P1: $\beta = 0.596$, $t = 65.48$, $p < .001$; P2: $\beta = 0.871$, $t = 338.7$, $p < .001$]).

There was also increased fixation of semantic competitors relative to unrelated distractors for all time windows in parameterisation 2 of the model (ts 1–16: $\beta = 0.148$, $t = 38.3$, $p < .001$; ts 17–32: $\beta = 0.964$, $t = 70.4$, $p < .001$; ts 33–48: $\beta = 0.969$, $t = 95.3$, $p < .001$; ts 49–64: $\beta = 0.874$, $t = 160.2$, $p < .001$) and for all time windows other than the first time window post word onset for parameterisation 1 of the model (ts 1–16: $\beta = -0.007$, $t = -1.32$, $p = .187$; ts 17–32: $\beta = 0.257$, $t = 13.26$, $p < .001$; ts 33–48: $\beta = 0.594$, $t = 42.68$, $p < .001$; ts 49–64: $\beta = 0.596$, $t = 41.98$, $p < .001$).

The rhyme effect increased above baseline levels for both parameterisations from the second time window post word onset and remained present for all remaining time windows (ts 1–16 [P1: $\beta = -0.002$, $t = -0.29$, $p = .774$; P2: $\beta = 0.006$, $t = 1.23$, $p = .218$]; ts 17–32 [P1: $\beta = 0.501$, $t = 20.94$, $p < .001$; P2: $\beta = 0.550$, $t = 47.6$, $p < .001$]; ts 33–48 [P1: $\beta = 1.620$, $t = 39.73$, $p < .001$; P2: $\beta = 1.100$, $t = 245.8$, $p < .001$]; ts 49–64 [P1: $\beta = 1.923$, $t = 47.72$, $p < .001$; P2: $\beta = 1.012$, $t = 113.0$, $p < .001$]).

As for the analysis of the MIM simulations, we examined whether the magnitude of competitor effects differed in each time window post word onset from the baseline time window pre-word onset using the same method as detailed in Section 'Simulation 2: Simulating effects of multimodal competition in the MIM – Results'. This analysis revealed that the log odds of fixating the visual competitor was greater than that of fixating the rhyme competitor in the first time window post word onset (ts 1–16: [P1: $\beta = 0.060$, $t = 7.23$, $p < .001$]; [P2: $\beta = 0.146$, $t = 23.9$, $p < .001$]) for both parameterisations of the model and in the second time window post word onset for the second parameterisation of the model (ts 17–32: [P1: $\beta = 0.034$, $t = 1.28$, $p = .201$]; [P2: $\beta = 0.421$, $t = 27.7$, $p < .001$]). However, in the third and fourth time window post word onset the log odds of fixating the rhyme competitor exceeded those of fixating the visual competitor for both parameterisations of the model (ts 33–48: [P1: $\beta = -0.932$, $t = -24.0$, $p < .001$]; [P2: $\beta = -0.128$, $t = -15.6$, $p < .001$]; ts 49–64: [P1: $\beta = -1.328$, $t = -30.1$, $p < .001$]; [P2: $\beta = -0.141$, $t = -14.0$, $p < .001$]).

Comparing fixations of the rhyme competitor to fixations of the semantic competitor revealed that for parameterisation 1 there was no difference from baseline levels in the first time window post word onset (ts 1–16: $\beta = -0.005$, $t = -1.06$, $p = .287$). However, fixation of rhyme competitors exceeded that of semantic competitors in all remaining time windows (ts 17–32: $\beta = -0.244$, $t = -10.6$, $p < .001$; ts 33–48: $\beta = -1.026$, $t = -20.8$, $p < .001$; ts 49–64: $\beta = -1.328$, $t = -25.5$, $p < .001$). For parameterisation 2 fixation of semantic competitors exceeded that of rhyme competitors in the first two time windows post word onset (ts 1–16: $\beta = 0.142$, $t = 45.5$, $p < .001$; ts 17–32: $\beta = 0.414$, $t = 27.1$, $p < .001$). However, fixation of rhyme competitors exceeded fixation of semantic competitors in the final two time windows (ts 33–48: $\beta = -0.131$, $t = -10.06$, $p < .001$; ts 49–64: $\beta = -0.138$, $t = -10.2$, $p < .001$).

For completeness, we also compared the log odds of fixating the visual competitor to the log odds of fixating the

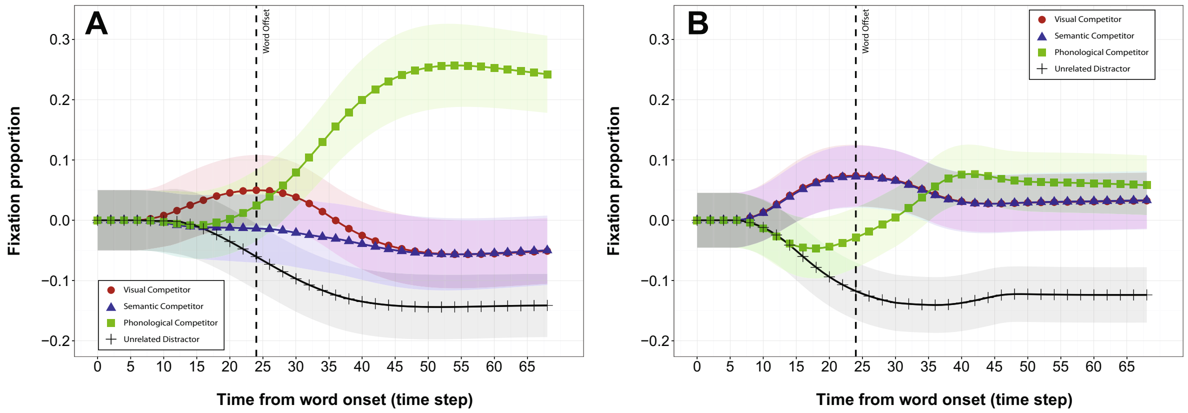

Fig. 6. Predictions generated by the extended TRACE model showing the change from word onset in the probability of fixating rhyme competitors, semantic competitors, visual competitors and unrelated distractors when all items are presented simultaneously in the same scene. (A) Behaviour generated by parameterisation 1, (B) behaviour generated by parameterisation 2. Shaded areas define 95% confidence intervals.

semantic competitor. For parameterisation 1 the visual competitor was fixated more than the semantic competitor in the first three windows post word onset [ts 1–16: $\beta = 0.065$, $t = 11.87$, $p < .001$; ts 17–32: $\beta = 0.278$, $t = 11.83$, $p < .001$; ts 33–48: $\beta = 0.094$, $t = 5.21$, $p < .001$], while there was no difference in the final time window [ts 49–64: $\beta = -0.0009$, $t = -0.008$, $p = .994$]. As expected, for parameterisation 2 there was no difference between the log odds of fixating the visual competitor and the log odds of fixating the semantic competitor for any time window [ts 1–16: $\beta = 0.004$, $t = 0.79$, $p = .430$; ts 17–32: $\beta = 0.006$, $t = 0.37$, $p = .715$; ts 33–48: $\beta = 0.003$, $t = 0.31$, $p = .757$; ts 49–64: $\beta = -0.002$, $t = -0.46$, $p = .647$].

Finally, using the same analysis technique described in Section 'Simulation 2: Simulating effects of multimodal competition in the MIM – Results' we examined whether the difference in empirical log odds of fixating the phonological rhyme competitor and the empirical log odds of fixating the unrelated distractor differed when visual and semantic competitors were either also present (simulation 4) or absent (simulation 3). Again we applied a linear mixed effects model with fixed effects of time window and scene (Simulation 3: rhyme competitor and unrelated distractors only; Simulation 4: rhyme competitor, semantic competitor, visual competitor and unrelated distractor) and random effects of item ($n = 10$) with random intercepts and slops for both time window and scene by item.

Analysing fixation behaviour in the first time window post word onset (ts 1–16) relative to the baseline (ts −6 to 0) window pre-word onset revealed no main effect ($t < 1.2$, $p > .25$), although there was a marginally significant interaction between Simulation and time window for parameterisation 2 ($\beta = 0.006$, $t = 1.653$, $p = .098$).

For parameterisation 1 of the model there was a significant main effect of Simulation and time and interaction between Simulation and time for all subsequent time windows (ts 17–32: time ∗ scene [$\beta = 0.074$, $t = 4.844$, $p < .001$], time [$\beta = 0.464$, $t = 18.99$, $p < .001$], scene [$\beta = 0.037$, $t = 3.425$, $p < .001$]; ts 33–48: time ∗ scene [$\beta = 0.311$, $t = 6.95$, $p < .001$], time [$\beta = 1.464$, $t = 43.56$, $p < .001$], scene [$\beta = 0.155$, $t = 4.92$, $p < .001$]; ts 49–64: time ∗ scene [$\beta = 0.428$, $t = 7.34$, $p < .001$], time [$\beta = 1.709$, $t = 57.54$, $p < .001$], scene [$\beta = 0.214$, $t = 5.54$, $p < .001$]). Examining beta estimates indicates that for parameterisation 1 the rhyme effect was greater post word onset, and was greater across all windows when semantic and visual competitors were also present in the display. Importantly the interaction between scene and time window indicated that the presence of visual and rhyme competitors increased the magnitude of the rhyme effect post word onset.

Analysing the results of simulations using parameterisation 2 also showed a significant interaction between time window and simulation for all subsequent time windows (ts 17–32: $\beta = 0.142$, $t = 2.188$, $p = .029$; ts 33–48: $\beta = 0.305$, $t = 2.670$, $p = .008$; ts 49–64: $\beta = 0.289$, $t = 2.868$, $p = .004$). There was also a main effect of time window for all subsequent windows (ts 17–32: $\beta = 0.479$, $t = 7.645$, $p < .001$; ts 33–48: $\beta = 0.948$, $t = 8.532$, $p < .001$; ts 49–64: $\beta = 0.867$, $t = 8.554$, $p < .001$), with no main effect of Simulation ($t < 1.50$, $p > .1$). Inspecting beta estimates indicate that there was an increase in the rhyme effect post word onset but importantly the TRACE model again predicts with this parameterisation an overall increase in the rhyme effect given the presence of visual and semantic competitors.

Summary

When feedback from semantic levels is delayed relative to visual level feedback (i.e., parameterisation 1) the onset of visual effects are predicted by TRACE to emerge shortly after word onset, with semantic effects emerging shortly after word offset and rhyme effects last to emerge. If the timing of cascading activation is equivalent for both visual and semantic activation then the extended TRACE model predicts that both visual and semantic effects should emerge shortly after word onset, with rhyme effects only emerging after word offset. This contrasts with predictions of the MIM system that predicts the emergence of all three competitor effects post word offset.

Both parameterisations of the extended TRACE system predict a greater influence of visual and semantic information on fixation behaviour, relative to phonological rhyme at early stages of word processing. However, unlike predictions of the MIM, at later stages of word processing (post word offset) the extended TRACE system predicts a greater influence of phonological rhyme overlap compared to overlap in semantic or visual dimensions.

A further point of distinction between model predictions is that for both parameterisations of the extended TRACE model, fixation of the rhyme competitor relative to the unrelated distractor is predicted to increase when visual and semantic competitors are present in the same display compared to when the rhyme competitor is presented only accompanied by unrelated distractors. The MIM by contrast predicted a reduced rhyme effect when additional competitors are present. These distinctions in prediction are consequences of the point at which information can integrate between multimodal representations. Note that the simulations in MIM and TRACE both provide broadly similar predictions about the role of onset phonological competitors in word processing, but differ over the extent to which rhyme competitors influence processing. We next test between these alternative accounts by experimental studies of the designs of Simulations 1 (3) and 2 (4).

Testing the effects of multimodal competition on phonological rhyme overlap in the visual world paradigm

The predictions of the two models were next tested in two visual world experiments that exposed participants to the same experimental conditions as were simulated.

Experiment 1: Effects of phonological rhyme overlap in target absent scenes

In Experiment 1 participants were presented with scenes containing four items while hearing a spoken target word. On experimental trials a single item within the display shared its phonological rhyme with the spoken target word and was the only relationship to exist between these two stimuli, with the remaining three items unrelated in visual, semantic and phonological dimensions.

Participants

40 participants (mean age = 21.6 years, range 18–30 years) recruited from the MPI for Psycholinguistics subject database were paid for participation in this study. All were native speakers of Dutch, had no known hearing problems and had corrected or normal vision.

Materials

15 experimental trials and 34 filler trails were constructed, each consisting of a visual display and spoken Dutch sentence. Each sentence consisted of a target word embedded in a neutral carrier sentence in which the target word was not predictable (e.g. Dutch: “Zij begrepen niet waarom de roos verwelkt was”, English Translation: “They could not understand why the rose was withered”). Approximately six words (Experimental trials: mean = 6.33, *SD* = 1.53; Filler Trials: 6.90, *SD* = 1.58) preceded the target word in the carrier sentences. Spoken sentences were recorded in a sound dampened room by a female native Dutch speaker who was not aware of the purpose of the study. Instruction was provided for sentences to be read in a neutral tone and to avoid highlighting individual words within the sentence.

Visual displays contained black and white line drawings of four objects. Each object was resized to fit an area 96 × 96 pixels. The four images were presented in the four corners of the 1024 × 768 pixel display (locations: 256 × 192; 256 × 576; 768 × 192; 768 × 578). Seventeen (target present) filler trials, two of which were used as practice trials, contained an image of the target word accompanied by three unrelated distractors. Seventeen (target absent) filler trials, two of which were used as practice trials, contained four unrelated distractors. Fifteen experimental trials contained a phonological rhyme competitor accompanied by three unrelated distractors (see Fig. 7).

Word frequency, number of letters and number of syllables were controlled between competitor and unrelated distractor sets (see Table 5). All target, phonological rhyme competitor and unrelated distractor words were monosyllabic (see Appendix Table C.1 and C.2 for a full list of experimental items). Phonological rhyme competitors were defined by the fact that they only differed from the target word in their initial phoneme (mean shared phonemes = 2.6, *SD* = 0.63). Controls ensured no sequence of phonemes was shared between target words and unrelated distractors. Separate semantic and visual similarity rating studies were conducted to ensure visual and semantic similarity was controlled across competitor and distractor sets.

Thirteen native Dutch speaking participants provided visual similarity ratings and a different group of 11 native Dutch speaking participants provided semantic similarity ratings, none of these participants later completed either of the eye tracking studies. Ratings were acquired using an online experiment in which participants were presented with the written form of the target word and the images corresponding to the rhyme competitor and distractors. In the case of visual similarity ratings participants were required to provide for each image a value between 0 and 10 indicating how similar the physical shape of the item in the image was to objects they associate with the target word (0 indicating no similarity in physical shape and 10 indicating both items have an identical physical shape), while ignoring any other relationships between the items for example semantic relationships. Similarly, for semantic similarity ratings, for each image participants provided a value between 0 and 10 indicating how much of the target words meaning is shared with the item depicted (0 indicating no similarity in meaning and 10 indicating complete overlap in meaning), while ignoring any other relationships between the items for example similarities in their physical shape. Results of these norming studies

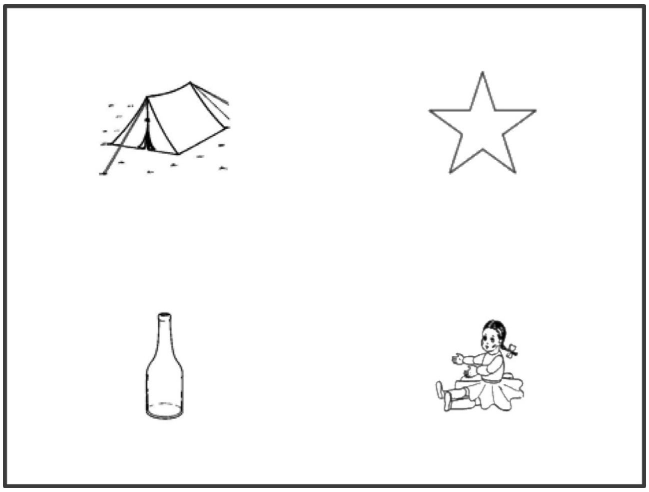

Fig. 7. Example of experimental display from Experiment 1. Within this trial the spoken target word was "cent", the rhyme competitor was 'tent' this is accompanied by three unrelated distractors 'pop' (doll), 'ster' (star) and 'fles' (bottle).

Table 5
Properties [μ (σ)] of words within competitor sets. Frequency = word frequency; Letters = number of letters; Syllables = number of syllables; Phonemes = number of phonemes shared with target spoken word; Semantic = semantic similarity rating to spoken target word; Visual = visual similarity rating to spoken target word.

Exp.	Item	Frequency	Letters	Syllables	Phonemes	Semantic	Visual
1	Rhyme	17.6 (23.24)	4.20 (0.41)	1.13 (0.35)	2.60 (0.63)	1.37 (0.46)	1.31 (0.84)
	Dist.1	27.8 (32.16)	4.73 (1.16)	1.27 (0.46)	0.33 (0.49)	1.36 (0.53)	1.33 (0.88)
	Dist.2	23.9 (33.88)	4.13 (0.35)	1.07 (0.26)	0.47 (0.83)	1.65 (0.78)	1.65 (0.97)
	Dist.3	35.0 (45.50)	4.27 (0.46)	1.27 (0.46)	0.33 (0.49)	1.41 (0.69)	1.42 (1.26)
2	Rhyme	17.6 (23.24)	4.20 (0.41)	1.13 (0.35)	2.60 (0.63)	0.95 (0.46)	1.23 (0.74)
	Sem.	21.5 (26.37)	4.67 (1.23)	1.33 (0.62)	0.47 (0.64)	5.90 (1.42)	2.30 (1.03)
	Visual	31.13 (81.82)	4.40 (0.91)	1.20 (0.41)	0.33 (0.62)	1.78 (0.89)	6.51 (1.13)
	Dist.	30.5 (32.06)	4.67 (1.23)	1.27 (0.46)	0.27 (0.46)	1.36 (0.67)	1.53 (0.90)

show that rhyme competitor and distractor sets did not differ in their semantic or visual similarity ratings to the spoken target words.

To ensure that the names attributed to displayed images were well motivated a picture name correspondence pre-test was conducted. 13 native Dutch speakers participated in this norming study and did not participate in either eye tracking experiment. Each image from experimental displays was presented to participants accompanied by either its intended name or a randomly selected name. Participants were required to indicate whether the name corresponded to the image or not. Of the 60 words tested 52 were rated as corresponding by 100% of participants, 6 words by 92%, 1 word (Dutch: *vest*; English: *waistcoat*) by 85% and 1 word (Dutch: *kennel;* English: *kennel*) by 75%.

Procedure

An Eyelink 1000 tower mounted eye tracker (sampling rate 1 kHz) was used to record participants' eye movements as they viewed displays on a computer monitor and listened to sentences through headphones while in a sound dampened room. Stimuli were presented and data recorded using the SR-Research program Experiment Builder.

Participants performed a 'look-and-listen' task (see Huettig, Olivers, & Hartsuiker, 2011, for further discussion), they were instructed to look at the screen while listening carefully to sentences they would hear through the headphones. Trials followed the same procedure as reported in Huettig and McQueen (2007). The experimenter initiated the start of each trial when the participant fixated a fixation cross in the centre of the screen, this allowed for drift correction in the calibration if required between trials. Once the trial was initiated the fixation cross remained in the centre of the screen for 500 ms, this was followed by a blank screen for 600 ms. Then a scene containing four images was presented with display onset coinciding with the onset of the spoken sentence. The scene remained displayed for 4300 ms (length of longest spoken sentence), following which a blank screen was presented for 500 ms after which the trial ended. Eye gaze was recorded at all stages of the trial. The location of targets and competitors was randomised across trials, while the location of items and order of trials was randomised across participants. Before the experiment began each participant first completed four practice trials (2 × target present filler trials, 2 × target absent filler trials). In total the experiment lasted approximately 15 min.

Results

Four interest areas were defined for each experimental display that covered the area 270 × 235 pixels that surrounded each image within the scene. A fixation was recorded as directed towards an item if it fell within the interest area within which the given item was situated. Blinks and saccades were not included in the analysis. Fig. 8 displays a time-course graph on which the difference in the proportion of fixations from target word onset directed towards rhyme competitors and the average unrelated distractor are plotted across the first 1600 ms post target word onset.

To examine the effect of the unfolding spoken target word on fixation behaviour we used a method of analysis similar to that used for analysing simulation results. For analysis we divided the first 1600 ms post target word onset into four 400 ms bins (1–400 ms; 401–800 ms; 801–1200 ms; 1201–1600 ms) and compared behaviour in each of these bins to behaviour in the 400 ms that preceded target word onset. For each bin in each trial we calculated the empirical log odds (see Jaeger, 2008) of fixating each category of item (i.e., rhyme competitor, unrelated distractor). The dependent measure was formed by subtracting the log-odds of fixating the unrelated distractor from the log-odds of fixating the phonological rhyme competitor. This difference measure reflects the difference in fixation behaviour as a consequence of phonological overlap. This measure in each of the 400 ms time windows post word onset was then compared to the 400 ms time window before word onset using linear mixed effect models to examine whether gaze was sensitive to phonological rhyme overlap in each of these post word onset periods. The model used to predict this variable applied the maximal random effect structure (Barr et al., 2013) with a fixed effect of window and random effects of subject and item. The random effects structure included random intercepts and slopes for time window both by subject and item. To derive *p*-values we assume *t*-values were drawn from a normal distribution (Barr, 2008).

A significant effect of phonological rhyme overlap was observed in the second time block (801–1200 ms) post word onset [$\beta = 0.68$; $t = 2.22$; $p = .03$]. The positive beta estimate indicates that phonological rhyme competitors were fixated above unrelated distractor levels in this time window. A marginal effect of phonological overlap was also observed in the third time block (1201–1600 ms) post word onset [$\beta = 0.05$; $t = 1.85$; $p = .06$]. There were no statistically robust effects of phonological rhyme overlap in any other time windows.

Discussion

The results of Experiment 1 demonstrated that systematic relationships embedded within the materials, specifically overlap in phonological rhyme shared between spoken target words and visually displayed phonological rhyme competitors, were sufficient to generate a phonological rhyme effect as has previously been described in visual world studies (Allopenna et al., 1998; McQueen & Huettig, 2012; McQueen & Viebahn, 2007), and similar in time course to that observed in Simulation 1 and 3 for both the MIM and extended TRACE model.

Experiment 2: Comparing phonological rhyme, visual and semantic overlap effects on language mediated visual attention

To test predictions of the two alternative models regarding the relative influence and timing of visual, semantic and phonological rhyme overlap effects on language mediated eye gaze participants gaze was recorded when viewing scenes containing a visual, a semantic and

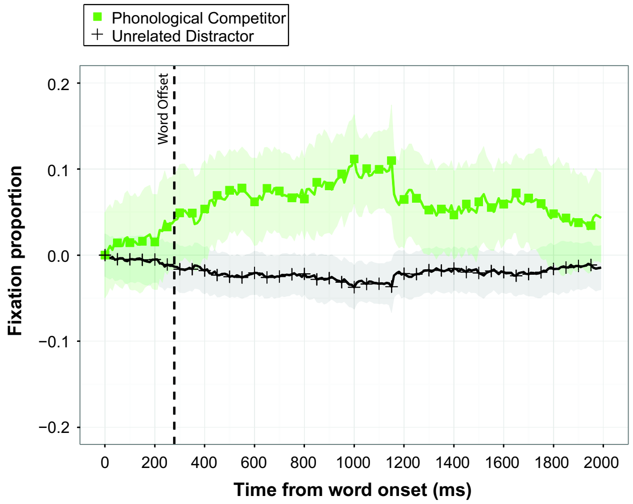

Fig. 8. Change in fixation proportions from target word onset in Experiment 1. Fixation proportions were averaged across all three unrelated distractors. Shaded areas define 95% confidence intervals.

a phonological rhyme competitor in addition to a single unrelated object.

Participants

39 participants (mean age = 25.3, range 18–30 years) took part in this study. All were recruited from the MPI for Psycholinguistics subject database and were paid for their participation. All participants were native Dutch speakers and had no known hearing problems and had normal or corrected to normal vision.

Materials

Experiment 2 used the same materials as used in Experiment 1, but with two of the distractors in experimental displays of Experiment 1 replaced by a visual competitor and a semantic competitor. Experiment 2 therefore also consisted of 15 experimental trials, 15 target absent filler trials and 15 target present filler trials. On experimental trials, scenes in Experiment 2 therefore now contained a phonological rhyme competitor, a visual competitor, a semantic competitor and an unrelated distractor (see Fig. 9).

All images of visual and semantic competitors were black and white line drawings and resized to fit the area 96 × 96 pixels. The four images were arranged evenly in four corners of the display using the same coordinates to centre objects as used in Experiment 1. Spoken sentences were also the same as those used in Experiment 1. Visual and semantic competitors were monosyllabic words and selected on the basis that they shared a visual or a semantic relationship with the spoken target word. Separate visual ($n = 13$) and semantic ($n = 10$) similarity norming studies were conducted to ensure only visual competitors differed from distractors in their level of visual similarity while only semantic competitors differed from distractors in levels of semantic similarity. Semantic competitors were rated marginally more visually similar to target words than unrelated distractors [$\mu = 0.77$; $\sigma = 1.41$ $p = .05$], while rhyme competitors were rated as marginally less semantically similar to target words than unrelated distractors [$\mu = -0.41$; $\sigma = 0.85$; $p = .08$]. It is likely that participants found it difficult to isolate the effects of visual or semantic similarity from overlap in other dimensions. Evidence for this can be found in the fact that that rhyme competitors were rated less semantically similar in Experiment 2 than in Experiment 1, even though participants were required to rate the same rhyme target combinations. Similarity ratings were collected from Dutch native speakers who did not participate in either eye tracking experiment using the same procedure outlined for norming of materials in Experiment 1. Further, phonological rhyme competitors were the only set to share an increased level of phonological overlap with the spoken target word. Competitor and distractor sets were also controlled for word frequency, number of letters and number of syllables (see Table 5).

Procedure

Experiment 2 followed a procedure identical to that described for Experiment 1 (see Section 'Experiment 1: Effects of phonological rhyme overlap in target absent scenes – Procedure').

Results

Fig. 10 displays the change in the proportion of fixations from target word onset directed towards each category of object (phonological rhyme competitor, visual competitor, semantic competitor, unrelated distractor) in Experiment 2 displays for the first 1600 ms post word onset.

Results of Experiment 2 were analysed using a similar method to that outlined for Experiment 1. However, in Experiment 2 there was a single distractor and three competitors. We therefore compared for each category of competitor (visual, semantic, rhyme) the difference between the empirical log odds of fixating a given competitor and the empirical log odds of fixating the distractor in the 400 ms prior to target word onset to the same measure

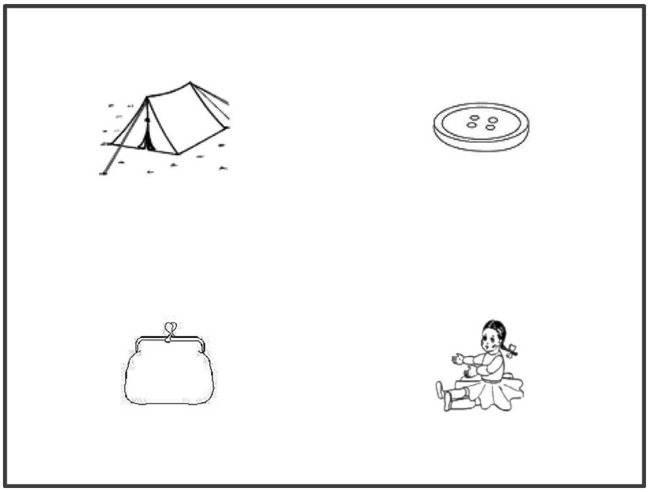

Fig. 9. Example of experimental display from Experiment 2. Within this trial the target word was "cent", the rhyme competitor was 'tent', the visual competitor 'knoop' (button), the semantic competitor 'beurs' (purse) and the unrelated distractor 'pop' (doll).

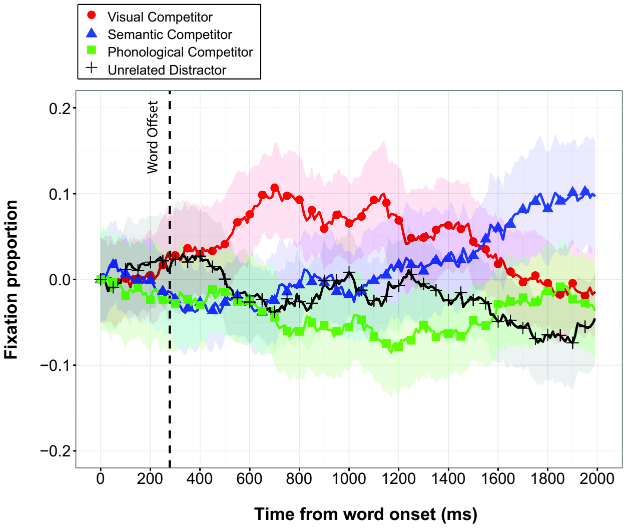

Fig. 10. Change in fixation proportions from target word onset directed towards rhyme competitors, visual competitors, semantic competitors and unrelated distractors in Experiment 2. Shaded areas define 95% confidence intervals.

calculated across one of four 400 ms time bins post word onset (1–400 ms; 401–800 ms; 801–1200 ms; 1201–1600 ms). This analysis revealed that visual competitors were fixated more than distractors in the second time block (401–800 ms) [$\beta = 0.67$; $t = 2.24$; $p = .03$], third time block [$\beta = 0.80$; $t = 2.68$; $p = .01$] and fourth time block [$\beta = 0.58$; $t = 2.01$; $p = .05$]. Semantic competitors were also fixated more than distractors although this effect emerged later, being marginally greater in the third time block [$\beta = 0.45$; $t = 1.79$; $p = .07$] and statistically robust in the fourth time block [$\beta = 0.66$; $t = 2.51$; $p = .01$]. There was, however, no evidence for an influence of phonological rhyme overlap on fixation behaviour as fixation of phonological rhyme competitors did not differ from distractors at any stage post word onset.

Finally, we examined whether the magnitude of competitor effects differed in each time window from the pre word onset baseline time window. This analysis revealed that the log odds of fixating the visual competitor did not differ from the log odds of fixating the semantic competitor in any time window post word onset [1–400: $\beta = -0.001$; $t = -0.004$; $p = .997$; 401–800: $\beta = 0.444$; $t = 1.48$; $p = .139$; 801–1200: $\beta = 0.349$; $t = 0.918$; $p = .359$; 1201–1600: $\beta = -0.076$; $t = -0.244$; $p = .807$]. The log odds of fixating the visual competitor were greater than the log odds of fixating the phonological rhyme competitor in time windows 401–800 ms [$\beta = 0.634$; $t = 2.14$; $p = .032$] and 801–1200 ms [$\beta = 0.906$; $t = 2.27$; $p = .023$], while differences were marginal in the final time window 1201–1600 ms [$\beta = 0.665$; $t = 1.74$; $p = .083$] and not significant in the initial window post word onset [1–400: $\beta = 0.103$; $t = 0.424$; $p = .672$]. The log odds of fixating the semantic competitor were greater than that of fixating the phonological rhyme competitor only in the final time window [1–400: $\beta = 0.104$; $t = 0.334$; $p = .738$; 401–800: $\beta = 0.190$; $t = 0.507$; $p = .612$; 801–1200: $\beta = 0.557$; $t = 1.549$; $p = .121$; 1201–1600: $\beta = 0.740$; $t = 2.48$; $p = .013$].

Discussion

Results of Experiment 2 show that visual properties shared between the spoken word and visually displayed items are first to bias attention followed by shared semantic properties. Semantic and visual similarity ratings suggest similar levels of overlap exist in the materials between competitor and distractor in both semantic and visual dimensions. This suggests that the initial bias towards visual distractors is driven by underlying architectural constraints of the system driving fixation behaviour or arises due to biases imposed by task specific constraints. For example, one explanation may be that visual information is prioritized when processing spoken words under the conditions imposed in this experiment as the task requires a mapping from a spoken word to an item's visual properties. This issue will be discussed further in the context of earlier simulation results in the General Discussion section of this paper. Although visual competitors were initially fixated more than semantic competitors, post hoc analysis shows a similar level of visual and semantic competitor fixation bias across the entire 1600 ms post word onset. This indicates that visual and semantic overlap exerts a similar level of influence on language mediated eye gaze. This is similar to a finding reported in Huettig and McQueen (2007) in which visual and semantic competitors were also presented to participants within the same scene. The pattern of results also suggests that the level of overlap implemented in the case of visual and semantic competitors was equivalent in the size of the elicited effect.

In contrast, however, although rhyme competitors only differed from target items in their initial phoneme and therefore overlapped significantly in a phonological dimension with the spoken target word, rhyme competitors failed to attract attention above control levels when visual and semantic competitors were also present. Previous visual world studies (Allopenna et al., 1998;

McQueen & Huettig, 2012; McQueen & Viebahn, 2007), including Experiment 1 reported in this paper, have demonstrated that visually displayed items that share their phonological rhyme with a spoken target word do bias fixation behaviour under conditions in which only a systematic phonological relationship exists between displayed items and spoken words. The results of Experiment 2, however, demonstrate that although the level of phonological rhyme overlap embedded in the materials is sufficient to influence eye gaze when phonological rhyme offers the only means of mapping between visually displayed items and spoken words (Experiment 1), this information does not exert an influence when semantic and visual information is also available to map between input streams. These data therefore show that visual and semantic relationships exert a greater influence on language mediated visual attention than phonological rhyme relationships to the extent that even when only a single phoneme in the phonological code mismatches there is no observable influence of phonological rhyme overlap on fixation behaviour. It should be noted, given recent work demonstrating the modulation of phonological rhyme influence by the level of noise in the speech signal (McQueen & Huettig, 2012), that these observed relative influences are likely to vary as a function of environmental factors such as quality of input signals. We debate this point further in the General Discussion.

Irrespective of the relative salience of phonological rhyme information in other conditions the combined data from Experiments 1 and 2 demonstrate the rapid activation of visual and semantic properties when processing spoken words. We know that the level of phonological rhyme overlap embedded in the materials is sufficient to generate an influence on fixation behaviour (Experiment 1). Therefore, for there to be no evidence for this effect in Experiment 2 the visual and semantic properties of the spoken target word must have been activated and available to map onto information activated by the visual display before overlapping phonological information in the rhyme of the word could begin to exert an influence on fixation behaviour.

Explicit awareness questionnaire

To assess participants' explicit awareness of the experimental manipulations within each experiment a short questionnaire was completed by participants once they had participated in either of the visual world experiments.

Participants

All participants in Experiment 1 ($n = 40$) and 2 ($n = 39$) completed the following questionnaire.

Materials & Procedure

Participants were asked to record on paper their response to the following questions: Heb je enige regelmaat at junnen ontdekken in de gerepresenteerde items? (English translation: Did you notice any relationships in the items presented?) to which they could respond 'Ja' (yes) or 'Nee' (No). If they responded 'Ja' then they were requested to provide a written description of the relationships they had noticed (Dutch Instruction: Zo ja, geef een beschruiving).

Table 6
Results of experimental manipulation awareness questionnaire.

	Exp. 1 ($n = 40$) Yes (%)	Exp. 2 ($n = 39$) Yes (%)
Express awareness	0.45	0.51
Identify Rhyme Competitors	0.05	0.08
Identify Semantic Competitors	0.03	0.21
Identify Visual Competitors	0.03	0.08

Results and discussion

Results of the questionnaire are summarised in Table 6. When scenes contained only rhyme competitors and unrelated distractors 21 of 40 participants indicated that they were not aware of any relationships between the items presented in the experiment. Of the 19 that indicated that they were aware of relationships between items only 2 participants explicitly recorded an awareness of a relationship between the sound of the words presented and items in the display. However, 1 participant recorded an awareness of items sharing a visual relationship, while another participant recorded an awareness of items sharing a relationship in their meaning even though neither semantic nor visual competitors were present.

In Experiment 2, when displays contained visual competitors, semantic competitors, phonological rhyme competitors and unrelated distractors 19 of 39 participants indicated that they were not aware of any relationships between the items presented in the experiment. Of the 19 that did indicate awareness, 3 indicated an awareness of a relationship between the sound of the word presented and items in the display. 3 participants also indicated an awareness of a visual similarity between items presented. 8 participants recorded an explicit awareness of a relationship in the meaning of the items presented.

The results of the questionnaire indicate that participants are largely unaware of the experimental manipulations within the materials. Although participants' gaze in Experiment 1 was sensitive to the overlap between the phonological rhyme of the spoken word and that corresponding to the phonological rhyme competitor, only 2 of 40 participants were able to indicate an explicit awareness of this sound similarity. Further, in Experiment 2 although robust visual and semantic competitor effects were observed, the vast majority of participants did not register an explicit awareness of similarities between the objects they viewed and words they heard in either of these modalities. The same number of individuals registered an awareness of visual similarity and sound similarity even though in Experiment 2 visual competitor effects were dominant and there was no evidence for sound similarity influencing fixations.

In order to examine whether eye gaze recorded in Experiments 1 or 2 was influenced by participants' expression of explicit awareness we performed additional post hoc tests. We ran the same mixed effect model analysis described in Sections 'Experiment 1: Effects of phonological rhyme overlap in target absent scenes – Results' and 'Experiment 2: Comparing phonological rhyme, visual and semantic overlap effects on language mediated visual attention – Results' but with an additional fixed effect of awareness. This analysis revealed no evidence for an effect of awareness on gaze behaviour critically with no significant interaction between time window and awareness in either Experiment 1 ($t < 0.5$, $p > .5$) or Experiment 2 ($t < 1.5$, $p > .1$), although the interaction was marginally significant for the visual effect in the time window 1200–1600 ms of Experiment 2 ($\beta = -1.05$, $t = 1.92$, $p = .055$). However, the negative estimate of this marginal effect suggests that should awareness have affected fixations of the visual competitor in this time window then the effect was to reduce the magnitude of the visual effect.

Taken together this suggests that the effects observed in both Experiments 1 and 2 represent early implicit processing of the concurrent visual and auditory stimuli that is likely to occur independently of participants' explicit goals, and therefore does not represent strategic processes explicitly engaged by participants given constraints imposed by the experimental manipulation.

General discussion

The purpose of this study was to determine the architectural constraints determining how multimodal information is integrated in speech processing. Together these models provide a first detailed comparison of how theories of multimodal information integration can be implemented in models of speech processing.

The two visual world studies examined the time course and relative influence of phonological rhyme, visual, and semantic information on language-mediated visual attention. The conditions to which participants were exposed in the two visual world studies were simulated in two models that represent alternative hypotheses for how multimodal information is integrated during language processing, the Multimodal Integration Model (MIM) and an extended multimodal TRACE model. MIM was chosen in order to generate predictions for how eye gaze is distributed given a fully interactive system and therefore will be referred to as the 'fully interactive' model, while the extended TRACE model was selected to generate predictions for how eye gaze is distributed given multimodal information is combined only at the lexical level and therefore will be referred to as the 'lexical level' model. We thus inferred properties of the underlying multimodal language processing system that is engaged when mapping between linguistic and non-linguistic information at the lexical level.

In the following sections we will first examine key features of the behavioural data, discussing their compatibility with each models' behaviour and the properties of the model that likely determine the ability or failure to replicate the behavioural results. We conclude by considering the broader implications for our understanding of the architecture supporting multimodal language processing.

Evaluating model predictions

We show that both the fully interactive architecture of the MIM and the lexical level interaction architecture of TRACE are able to generate behaviour consistent with previous visual world studies that have investigated the influence of visual, semantic and onset phonological information on gaze behaviour. Critically however, unlike previous data sets, the novel experimental data presented in this paper distinguishes between models, with models differing in their prediction of the influence of phonological rhyme relative to semantic and visual information post word offset.

Our experimental results showed that under conditions in which visual, semantic and phonological rhyme information are all available to constrain word referent mapping, visual and semantic relationships dominate such that phonological rhyme exerts no observable influence on behaviour. The data thus demonstrate that visual and semantic information is activated rapidly by the incoming speech signal and can be recruited by the cognitive system to map onto pre-activated information (i.e., activated via the immediate visual environment). The speed with which such multimodal activation and integration occurs in Experiment 2 is sufficient to ensure that information activated by later phonemes in the rhyme of a word exert no observable influence on behaviour, even though Experiment 1 demonstrates that sufficiently strong relationships exist within the materials to generate detectable effects when competing visual and semantic information is not present.

In both the MIM and TRACE models, visual and semantic information is activated rapidly and is available to constrain fixation behaviour shortly after word onset. However, the two models make distinct predictions regarding the influence of phonological rhyme information on fixation behaviour post word offset. Although the fully interactive system predicts that phonological rhyme competitors will be fixated more than unrelated items when visual and semantic competitors are also present, rhyme competitors are predicted to be fixated at substantially lower levels than visual or semantic competitors. Therefore, consistent with the behavioural results the fully interactive system predicts a consistently greater influence post word offset of shared semantic and visual properties on fixation behaviour than shared phonological properties in the rhyme. Further, the fully interactive system of the MIM predicted that the presence of semantic and visual competitors would reduce the phonological rhyme effect, a property that is also consistent with observed behaviour. Though much reduced, in the MIM the level to which the rhyme effect was reduced was not sufficient to entirely eliminate the effect as was observed in the behavioural data.

The TRACE model also predicts that, in a system with multimodal integration at the lexical level, visual competitors and semantic competitors (given rapid cascading of activation to semantic levels, Parameterisation 2) will be fixated more than phonological rhyme competitors at early stages of spoken word processing, however in contrast to the behaviour observed, at later stages of processing phonological rhyme competitors attract increased fixation, to the extent that they are fixated above the level of visual and semantic competitors. The lexical level model thus appears to overestimate the relative influence of phonological rhyme information on gaze behaviour. Further, when comparing predicted behaviour across experimental conditions the lexical level interaction in the TRACE model predicted an increase in the magnitude of the phonological rhyme effect when rhyme competitors are presented in the same scene as semantic and visual competitors. This also contrasted with the behavioural data in which rhyme effects are reduced to the extent that they no longer exert an effect on behaviour when visual and semantic competitors are present.

As has been observed in previous visual world studies (Huettig & McQueen, 2007), distinctions in the behavioural data between visual and semantic effects were less clear than distinctions between visual (or semantic) and phonological effects. Visual effects were first to emerge shortly after word onset, while semantic effects did not emerge until time windows post word offset. However, although beta estimates, which we interpret as estimates of effect size, were numerically higher in early windows for visual competitors and higher for semantic competitors in later time windows, when analysing differences between fixation of semantic and visual competitors directly no significant difference was revealed in any time window of analysis. Simulations with the fully interactive model (MIM) did not predict the observed difference in the onset of visual and semantic effects, yet, as in the behavioural data, direct comparisons of fixations of visual and semantic competitors generated by the model did not reveal a significant difference. The lexical level interaction system of the TRACE model generated differences between visual and semantic effects only when the cascading of activation from semantic levels was delayed relative to visual levels. The implemented delay led to a replication of the observed earlier onset of visual effects. This also however generated significantly greater visual effects in the first three time windows, a feature not present in the behavioural data, although by the final time window this visual advantage was no longer present.

Determining the mechanisms of the MIM that drive observed effects

We have thus far considered that the ability of the MIM to replicate both the increased visual and semantic effect relative to phonological rhyme and the reduced rhyme effect in the presence of visual and semantic competitors is due to the fully interactive integration resulting from the model's architecture. However, other features of the MIM may also have contributed to the qualitative effects of different representation types. There is a potential (additional) role of the properties of its representations, in particular the sequential phonological input relative to the simultaneous availability of visual and semantic information. Alternatively, effects may instead be a consequence of the MIM's training regime, given that error signals for certain representations were presented at different temporal points during training trials, and certain mappings were trained with greater frequency than others. In order to identify which properties of the MIM influenced which properties of its behaviour we ran a series of post hoc simulations.

Phonological representations within the MIM are unlike visual or semantic representations in that they unfolded over time with an additional phoneme presented at each time step. This property of the model's design was chosen in order to simulate the fact that unlike the information in the visual display which becomes available at a single time point, information in the auditory signal becomes available at different points in time as the speech signal unfolds. We removed this aspect of the phonological representations by presenting to the MIM the full phonological form of the target word at word onset both in training and testing. Following this manipulation, as expected, no significant difference was observed between the effect of phonological onset and phonological rhyme overlap on gaze behaviour displayed by the MIM (see Appendix A: MIM simulation Ph. 1,a). However, comparisons between the MIM trained with or without a temporal component to the phonological representations showed no significant difference in the magnitude of the rhyme effect, while rhyme competitors were still fixated below the level of visual and semantic distractors when presented in the same scene (see Appendix A: MIM simulation Ph. 1,b). This indicates that the temporal component of phonological representations, although generating important distinctions between the effect of phonological rhyme and phonological onset overlap, did not drive the greater visual and semantic effects relative to phonological rhyme in the MIM system. As the TRACE model also incorporated sequential phonological presentations, the source of the relative effects of phonological rhyme, semantic, and visual information in the MIM must result from other aspects of the model's architecture or training.

A second feature of the MIM that defines differences between the processing of visual, semantic and phonological information, and thus may drive distinctions in the effects observed, is the onset of the training signal in training trials. Due to the temporal component of phonological representations, within MIM time is required to elapse before sufficient information is available in the auditory input to identify the target. By contrast when a visual input or semantic representation is presented to MIM there is immediately sufficient information in the signal to identify the target. For this reason, the training signal is provided later for phonology to semantic mappings and phonology and vision to location mappings, as opposed to vision to semantic and semantic and vision to location mappings. We therefore ran an additional set of post hoc simulations in which the MIM was trained with the onset of the training signal equated across all training tasks (i.e. word offset). This change to MIM training did not alter the magnitude of competitor effects observed and thus it was not sufficient to explain increased fixations to visual and semantic competitors over phonological rhyme competitors (see Appendix A: MIM simulation Ph. 2).

The final property of MIM training that may influence simulated differences between competitor effects may be found in differences in the proportion of training on each task, with training on phonology driven orientation four times less likely to occur than other training tasks. This initial design decision was motivated by evidence that in the natural learning environment items surrounding a child are frequently left unnamed (Yu & Ballard, 2007). However, we removed this assumption from the model to examine its effects on behaviour (see Appendix A: MIM simulation Ph. 3). This resulted in a marginal reduction in the semantic effect relative to the visual effect in the MIM. This suggests that without additional training on semantic orientation tasks the system is more dependent on direct visual associations than indirect associations that may be learnt between phonological and visual properties via semantics.

Together these post hoc simulations have identified properties of MIM that allow it to replicate unique properties of the time course of visual, semantic, and phonological rhyme and onset competitor effects. These additional simulations indicate that the larger effect of the visual competitors compared to semantic and phonological competitors, observed in the MIM simulations likely reflect the manner in which the system is exposed to visual information during training. Visual information is essential to both semantic and phonologically driven orientation tasks, so visual representations must be isolated and prioritised in a multi-object input. This likely resulted in an enhancement of the visual effects in the MIM's performance.

Assumptions in extending the TRACE model to multimodal stimuli

In terms of the role of interactivity in producing the observed multimodal integration effects, these effects can be assessed by analysing the performance of the TRACE model, where no sub-lexical integration is permitted. The predictions of eye gaze behaviour from two parameterisations of the extended TRACE system both overestimated the phonological rhyme effect post word offset. We believe this to be a property consistent with lexical level alignment models, such as those represented by the extended TRACE system, however, our extension of the TRACE model required a set of assumptions to be implemented. We assumed that activation from lexical level nodes cascades to activate associated properties at visual and semantic levels in TRACE, to the extent that rhyme competitors' visual and semantic properties are activated to at least 5% the level of the target's visual and semantic properties should it be present in the visual display (Post hoc simulations indicate that rhyme effects still exceed or equal semantic and visual effects post word offset with rhyme competitors provided 2.5% target level activation; see Appendix B: TRACE simulation Ph. 1.a and Ph. 1.b). This assumption was based on the results of Allopenna et al. (1998) that show rhyme competitors are activated at the lexical level at approximately 10% of the target word's level when the target is present in the display (therefore the implementation of 5% represents a conservative estimate). One method for reducing the level of predicted phonological rhyme effects from below their current over-estimated high level in the TRACE simulations, would be to decrease the level of activation cascading from the lexical level to activate visual and semantic properties relating to the rhyme competitor. However, although this may improve the fit with observations in Experiment 2 of this study it would compromise the TRACE model's ability to generate rhyme effects given the presence of a target item and phonological onset competitor in the display (as in Allopenna et al., 1998).

Alternatively, the relative levels of fixation of semantic, visual and phonological rhyme competitors predicted by TRACE could be altered by increasing fixation to visual and semantic competitors. This could be achieved either by increasing the level of activation cascading to activate lexical level nodes from visual and semantic levels for visual and semantic competitors, respectively, or by increasing the scaling factor determining the influence of such activation on lexical level node activation. We currently assume that visual and semantic competitors benefit from activation cascading from visual and semantic levels respectively at 50% the level of the target item (when a target is present in the display). We believe a value of 50% is justified as it closely approximates the level of similarity between target and competitor judged by participants in independent visual and semantic similarity rating studies (see Huettig & McQueen, 2007; Table 5 in this paper). However, additional simulations of TRACE indicated that fixation of rhyme competitors would still exceed or equal that of visual and semantic competitors post word onset even if visual and semantic activation was 100% target levels for visual and semantic competitors respectively (see Appendix B: TRACE simulation Ph. 2.a and Ph. 2.b).

In further simulations, we have also explored the effects of increasing visual and semantic scaling factors (s_f, s_p) in the extended TRACE model, however such increases lead to distortions of the smooth time course fixation functions particularly in the critical period post word offset. Further, post hoc simulations indicate that doubling the scaling factor $s_{f,p} = 1$ still leads to fixation of phonological rhyme competitors at levels similar to visual and semantic competitors post word offset (irrespective of whether the lexical level node corresponding to the target also receives activation from visual and semantic levels [P2] or not [P1]; see Appendix B: TRACE simulation Ph. 3.a and Ph. 3.b).

Conclusions and consequences for models of (multimodal) language processing

The fully interactive (MIM) and lexical level (TRACE) models make contrasting predictions regarding how the presence or absence of visual and semantic competitors will impact the overall magnitude of the phonological rhyme effect. The MIM correctly predicted that the inclusion of visual and semantic competitors would lead to a decrease in the magnitude of the phonological rhyme effect, while both parameterisations of the lexical level system predicted an increase. This suggests that within MIM greater activation of properties associated with the target, i.e. its visual and semantic features, leads to increased inhibition of the influence of properties associated with the phonological rhyme of the word but not the target. By contrast as multimodal interaction is limited to the lexical level with our extended TRACE system such complex sub-lexical cross modal associations cannot develop to exert such an effect.

There still remain however, two features of the data that the MIM failed to fully replicate. These were that the MIM did not entirely eliminate the rhyme effect when presented alongside visual and semantic competitors, and also in the behavioural data fixation of visual competitors increased above unrelated distractor levels earlier than semantic competitors while in MIM fixation of visual and semantic competitors departed from unrelated distractor levels in the same time window. However, such observations are not incompatible with the MIM framework. In the case of elimination of the rhyme effect, previous simulations with the MIM have demonstrated that exposure to noise in the phonological signal during training was critical to generating phonological rhyme effects (Smith et al., 2013). Reducing the amount of noise to which the system is exposed during training is therefore likely to reduce the magnitude of the phonological rhyme effect. Given that the presence of visual and semantic competitors was shown to reduce the phonological rhyme effect, it is therefore feasible that a level of noise could be found that would give rise to a phonological rhyme effect when placed in scenes with only unrelated distractors, yet would not be observable when presented alongside visual and semantic competitors.

Concerning the temporal visual and semantic effects, we have observed through the post hoc simulations reported above that the amount of training on semantic mapping tasks relative to phonological mapping tasks alters the magnitude of the semantic competitor effect relative to the visual competitor effect. It is therefore feasible that reducing the semantic bias during training, and thus the magnitude of the semantic effect relative to the visual effect, will result in an earlier onset of visual effects. Importantly, however, although the onset of visual effects was observed to be earlier than semantics in this study, there was no significant difference between the magnitude of visual and semantic fixations in any time windows examined in the MIM. Further, in Huettig and McQueen (2007) differences in fixation of visual and semantic competitors were not observed when visual displays were previewed for an extended period of 1 s prior to word onset, which was a feature of the experimental design that was implemented in the experiments reported in this paper. It appears therefore that distinctions in the onset of visual effects relative to semantic are likely to be small or marginal under such long preview conditions.

The above evaluation of models against the behavioural data demonstrates that only the fully interactive architecture is compatible with the behavioural results of Experiment 2 of this study. Attempting to tease apart the factors that generate the emergent properties of such a fully interactive multimodal system is complex. However, the MIM's success in replicating the complex, temporal effects of phonology, semantics and vision appears due to the interactivity of information sources at all stages of processing mappings between multiple modalities facilitated by its architecture. By contrast the simulations conducted using the extended multimodal TRACE model indicate that a system that restricts interaction between modalities to the lexical level is likely to over emphasise the influence of phonological rhyme information on processing.

The experimental data presented in this paper demonstrates that phonological rhyme information exerts little or no influence on language mediated eye gaze in contexts in which visual and semantic information is also available to influence fixations. Previous studies of language mediated eye gaze have been weakened for frequently lacking an explicit description of the connection between the indirect measure of gaze and the underlying cognitive processes that it is argued to represent (see Anderson, Chiu, Huette, & Spivey, 2011; Ferreira & Tanenhaus, 2007; Huettig et al., 2011; for discussion). In this paper we describe explicitly two alternative architectures capable of supporting the multimodal processes required to generate language mediated eye gaze at the level of single words. We demonstrate that both a fully interactive system and a system in which multimodal information interacts only at the lexical level are able to generate behaviour consistent with previous language mediated eye gaze data sets. However, in a novel visual world study we demonstrate it is possible to tease apart these alternative architectures by examining their predictions for the influence of phonological rhyme relative to semantic and visual information. Contrary to observed behaviour the lexical level system predicted a greater influence of rhyme information post word onset, whereas the fully interactive system correctly predicted a greater influence of visual and semantic information.

We acknowledge that language mediated visual attention only offers an indirect measure of the processes underlying spoken word recognition and comprehension. However, a fundamental property of language processing is its ability to connect information across modalities. Furthermore, given the ambiguities present in natural language (Ferreira, 2008; Jaeger, 2006, 2010; Piantadosi et al., 2012; Roland et al., 2006; Wasow & Arnold, 2003; Wasow et al., 2005) should the spoken word recognition system have access to multimodal information, an efficient spoken word recognition system should rapidly accommodate such cues adapting its response in accordance to the current multimodal evidential landscape. A comprehensive description of the language processing system should therefore describe the architecture that supports such multimodal interaction.

We believe that language mediated visual attention can provide important clues as to the nature of the architecture supporting such multimodal interaction as it captures changes in behaviour as participants are required to map between linguistic and non-linguistic information. The results of our post-experiment questionnaire indicate that the effects captured by the visual world experiments in this paper reflect early implicit processing that is likely to occur independent of a participant's explicit awareness or goals.

Taken together, the results of the visual world simulations and experiments conducted in this study clearly show that visual and semantic information is activated rapidly and is available to constrain behaviour as a spoken word unfolds. Therefore, the cognitive system has access to multimodal information in which reliable cues as to the meaning of a given utterance are likely to exist and thus it seems probable that an efficient system it is likely to make use of this information to constrain spoken word recognition. Models of speech recognition have frequently overlooked the multimodal nature of the speech recognition problem in "real world" environments. Most past studies have focussed purely on the phonological properties of the system (e.g. Luce et al., 2000; McClelland & Elman, 1986; Norris & McQueen, 2008; Scharenborg & Boves, 2010) which we believe is likely to describe only a single component of a complex multimodal system.

We have here made some initial steps towards defining such a comprehensive model by testing two alternative architectures that describe explicitly the points of processing at which linguistic and non-linguistic information may interact during spoken word processing. Each model describes language processing in terms of multimodal constraint satisfaction, allowing visual and semantic information to rapidly constrain language processing within a rapidly cascading (see Pulvermüller et al., 2009) or parallel architecture. However, the models differed in relation to the level of representation at which information is able to interact across modalities. Our results suggest that multimodal interaction during language processing is not restricted to the lexical level but instead supported by an architecture that facilitates interaction of information across modalities at both lexical and sub-lexical levels of representation.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jml.2016.08.005.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*(4), 419–439.

Anderson, S. E., Chiu, E., Huette, S., & Spivey, M. J. (2011). On the temporal dynamics of language-mediated vision and vision-mediated language. *Acta Psychologica, 137*(2), 181–189.

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition, 52*(3), 163–187.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*(4), 457–474.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Connine, C. M., Blasko, D. G., & Titone, D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition? *Journal of Memory and Language, 32*(2), 193–210.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology, 6*, 84–107.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology, 42*(4), 317–367.

Dahan, D., & Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic Bulletin & Review, 12*(3), 453–459.

Dilkina, K., McClelland, J. L., & Plaut, D. C. (2008). A single-system account of semantic and lexical deficits in five semantic dementia patients. *Cognitive Neuropsychology, 25*(2), 136–164.

Dilkina, K., McClelland, J. L., & Plaut, D. C. (2010). Are there mental lexicons? The role of semantics in lexical decision. *Brain Research, 1365*, 66–81.

Duñabeitia, J. A., Avilés, A., Afonso, O., Scheepers, C., & Carreiras, M. (2009). Qualitative differences in the representation of abstract versus concrete words: Evidence from the visual-world paradigm. *Cognition, 110*(2), 284–292.

Ferreira, V. (2008). Ambiguity, accessibility, and a division of labor for communicative success. *Psychology of Learning and Motivation, 49*, 209–246.

Ferreira, F., & Tanenhaus, M. K. (2007). Introduction to the special issue on language–vision interactions. *Journal of Memory and Language, 57*(4), 455–459.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences, 6*(2), 78–84.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12*(5–6), 613–656.

Hockett, C. F., & Altmann, S. (1968). A note on design features. In T. A. Sebeok (Ed.), *Animal communication: Techniques of study and results of research* (pp. 61–72). Bloomington: Indiana University Press.

Huettig, F., & Altmann, G. T. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition, 96*(1), B23–B32.

Huettig, F., & Altmann, G. T. (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. *Visual Cognition, 15*(8), 985–1018.

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language, 57*(4), 460–482.

Huettig, F., Mishra, R. K., & Olivers, C. N. (2012). Mechanisms and representations of language-mediated visual attention. *Frontiers in Psychology, 2*, 394.

Huettig, F., Olivers, C. N., & Hartsuiker, R. J. (2011). Looking, language, and memory: Bridging research from the visual world and visual search paradigms. *Acta Psychologica, 137*(2), 138–150.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137*(2), 151–171.

Jaeger, T. F. (2006). *Redundancy and syntactic reduction in spontaneous speech. Unpublished doctoral dissertation*. Stanford University.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434–446.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology, 61*, 23–62.

Kurumada, C., & Jaeger, T. F. (2015). Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language, 83*, 152–178.

Leonard, M. K., & Chang, E. F. (2014). Dynamic speech representations in the human temporal lobe. *Trends in Cognitive Sciences, 18*(9), 472–479.

Luce, R. D. (1959). *Individual choice behaviour: A theoretical analysis*. New York: Wiley.

Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception & Psychophysics, 62*(3), 615–625.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*(4), 676.

Magnuson, J. S., Tanenhaus, M. K., & Aslin, R. N. (2000). Simple recurrent networks and competition effects in spoken word recognition. *University of Rochester Working Papers in Language Science, 1*, 56–71.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (2003). The time course of spoken word learning and recognition: Studies with artificial lexicons. *Journal of Experimental Psychology: General, 132*(2), 202–227.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*(1), 71–102.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*(1), 1–86.

McClelland, J. L., Mirman, D., Bolger, D. J., & Khaitan, P. (2014). Interactive activation and mutual constraint satisfaction in perception and cognition. *Cognitive Science, 38*(6), 1139–1189.

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences, 10*(8), 363–369.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The appeal of parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, & the PDP research group (Eds.). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. I, pp. 3–44). Cambridge, MA: MIT Press.

McQueen, J. M., & Huettig, F. (2012). Changing only the probability that spoken words will be distorted changes how they are recognized. *The Journal of the Acoustical Society of America, 131*(1), 509–517.

McQueen, J. M., Norris, D., & Cutler, A. (2006). Are there really interactive processes in speech perception? *Trends in Cognitive Sciences, 10*(12), 533.

McQueen, J. M., & Viebahn, M. C. (2007). Tracking recognition of spoken words by tracking looks to printed words. *The Quarterly Journal of Experimental Psychology, 60*(5), 661–671.

Mirman, D., & Magnuson, J. S. (2009). Dynamics of activation of semantically similar concepts during spoken word recognition. *Memory & Cognition, 37*(7), 1026–1039.

Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General, 140*(3), 325.

Monaghan, P., & Nazir, T. (2009). Modelling sensory integration and embodied cognition in a model of word recognition. *Connectionist Models of Behaviour and Cognition II*, 337–348.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review, 115*(2), 357.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52*(3), 189–234.

Onnis, L., & Spivey, M. J. (2012). Toward a new scientific visualization for the language sciences. *Information, 3*(1), 124–150.

Pearlmutter, B. A. (1989). Learning state space trajectories in recurrent neural networks. *Neural Computation, 1*(2), 263–269.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition, 122*(3), 280–291.

Plaut, D. C. (2002). Graded modality-specific specialisation in semantics: A computational account of optic aphasia. *Cognitive Neuropsychology, 19*(7), 603–639.

Pulvermüller, F., Shtyrov, Y., & Hauk, O. (2009). Understanding in an instant: Neurophysiological evidence for mechanistic language circuits in the brain. *Brain and Language, 110*(2), 81–94.

R Development Core Team (2009). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL <http://www.R-project.org/>.

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). Structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review, 111*(1), 205.

Rogers, T. T., & McClelland, J. L. (2014). Parallel distributed processing at 25: Further explorations in the microstructure of cognition. *Cognitive Science, 38*(6), 1024–1077.

Roland, D., Elman, J. L., & Ferreira, V. S. (2006). Why is that? Structural prediction and ambiguity resolution in a very large corpus of English sentences. *Cognition, 98*(3), 245–272.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. *Volume I: Foundations & volume II: Psychological and biological models*. Cambridge, MA: MIT Press.

Scharenborg, O., & Boves, L. (2010). Computational modelling of spoken-word recognition processes: Design choices and evaluation. *Pragmatics & Cognition, 18*(1), 136–164.

Smith, A., Monaghan, P., & Huettig, F. (2013). An amodal shared resource model of language-mediated visual attention. *Frontiers in Psychology, 4*, 528.

Smith, A. C., Monaghan, P., & Huettig, F. (2014a). Literacy effects on language and vision: Emergent effects from an amodal shared resource (ASR) computational model. *Cognitive Psychology, 75*, 28–54.

Smith, A. C., Monaghan, P., & Huettig, F. (2014b). Modelling language – Vision interactions in the hub and spoke framework. In J. Mayor & P. Gomez (Eds.), *Computational models of cognitive processes: Proceedings of the 13th neural computation and psychology workshop (NCPW13)* (pp. 3–16). Singapore: World Scientific Publishing.

Spivey, M. (2007). *The continuity of mind*. Oxford: Oxford University Press.

Strauss, T. J., Harris, H. D., & Magnuson, J. S. (2007). JTRACE: A reimplementation and extension of the TRACE model of speech perception and spoken word recognition. *Behavior Research Methods, 39*(1), 19–30.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*(5217), 1632–1634.

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research, 29*(6), 557–580.

Wasow, T., & Arnold, J. (2003). Post-verbal constituent ordering in English. *Determinants of Grammatical Variation in English*, 119–154.

Wasow, T., Perfors, A., & Beaver, D. (2005). The puzzle of ambiguity. *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe*. Stanford: CSLI Publications.

Yee, E., Huffstetler, S., & Thompson-Schill, S. L. (2011). Function follows form: Activation of shape and function features during object identification. *Journal of Experimental Psychology: General, 140*(3), 348–363.

Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(1), 1–14.

Yee, E., Overton, E., & Thompson-Schill, S. L. (2009). Looking for meaning: Eye movements are sensitive to overlapping semantic features, not association. *Psychonomic Bulletin & Review, 16*(5), 869–874.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing, 70*(13), 2149–2165.